Die Sicht von Auftraggebern auf den Evaluationsprozess

Unabhängigkeit von Evaluationen und Einflussnahme im Evaluationsprozess aus der Perspektive von Auftraggebern in den USA

Autorin:	Fabienne Helen Schmidli
	schmifa7@students.zhaw.ch 13-611-348
Hauptbetreuung:	Dr. Lyn Pleger
Co-Betreuung:	Susanne Hadorn, Kompetenzzentrum für Public Management, Universität Bern
Studiengang:	Master in Business Administration with a Specialization in Public and Nonprofit Management

Schriftliche Arbeit verfasst an der School of Management and Law,
Zürcher Hochschule für angewandte Wissenschaften

Winterthur, 14. Juni 2019

Management Summary

Im Evaluationskontext gilt die wissenschaftliche Unabhängigkeit als wesentliche Voraussetzung für Evaluationen. Dennoch stehen Evaluierende im Spannungsfeld die Unabhängigkeit von Evaluationen aufrechtzuerhalten und gleichzeitig den Erwartungen der Auftraggeber gerecht zu werden. Auftraggeber von Evaluationen gelten als diejenigen Stakeholder mit der grössten Einflussnahme im Evaluationsprozess. Dabei kann ihre Druckausübung viele Formen annehmen, wobei die für die Unabhängigkeit von Evaluationen wichtigen Faktoren der Objektivität und wissenschaftlichen Integrität gefährdet werden. Während sich bisherige Forschung auf der Perspektive von Evaluierenden und ihrem Umgang mit ethischen Herausforderungen fokussierte, herrscht zur Auftraggeberseite eine Forschungslücke.

Die Masterarbeit gilt als erster Versuch einen umfassenden Überblick über die Auftraggeberperspektive der US-Evaluationslandschaft zu schaffen. Konkret wird untersucht, wie Auftraggeber die Unabhängigkeit von Evaluationen beurteilen und welche Rolle die Eigenschaften der Auftraggeber und ihre Beziehung zu Evaluierenden in der Beeinflussung des Evaluationsprozesses spielen. Dabei wird die Principal-Agent-Theorie im Kontext von Evaluationen validiert, wobei der Fokus auf den Auftraggebern als Principals liegt, welche die Evaluierenden (Agents) in ihrer Evaluationstätigkeit beeinflussen. Dazu wurde eine Online-Befragung von Auftraggebern in den USA durchgeführt, die zusätzlich durch die American Evaluation Association unterstützt wurde. Ergänzend zur Hypothesenüberprüfung wurde zur Validitätsprüfung der Befunde der Schweizer Vorbildstudie von Pleger und Hadorn (2018) ein Ländervergleich durchgeführt, wobei Gemeinsamkeiten und Unterschiede identifiziert wurden.

Die Befunde zeigen, dass die Bandbreite möglicher Beeinflussungsformen – ob konstruktiv oder destruktiv – sehr gross ist. Generell wurde ein positiver, statistisch signifikanter Zusammenhang sowohl zwischen der Unzufriedenheit und der destruktiven Einflussnahme als auch dem Konfliktverhältnis und dem Anreizsystem identifiziert. Für die konstruktive Einflussnahme auf den Evaluationsprozess wurden positive signifikante Zusammenhänge mit der Vertrautheit als auch der Berufserfahrung gefunden. Analog zur Schweizer Studie überrascht für den US-Evaluationskontext, dass zwar generell ein kon-

fliktgeprägtes Verhältnis vorherrscht, jedoch keinem Auftraggeber jemals eine Druckausübung unterstellt wurde. Zudem wurde ein alarmierendes Informationsdefizit aufseiten der Auftraggeber hinsichtlich der Bekanntheit der nationalen „Program Evaluation Standards“ identifiziert.

Zur Bereicherung des gegenseitigen Austauschs zwischen Auftraggebern und Evaluierenden sowie der Gewährleistung der Unabhängigkeit von Evaluationen werden drei Handlungsempfehlungen abgeleitet. Erstens die Schaffung eines grösseren, gegenseitigen Verständnisses zwischen beiden Akteuren. Zweitens eine offene Kommunikationskultur zur Konfliktlösung und -prävention und drittens, breitangelegte Informationskampagnen der Evaluationsgemeinschaft zur Bekanntheitssteigerung von Evaluationsstandards.

Inhaltsverzeichnis

Abkürzungsverzeichnis

AEA	American Evaluation Association
BUSD	Betterment, Undermining, Support and Distortion
EBP	Evidence-based policy making
NGO	Nichtregierungsorganisation
NPO	Nonprofit-Organisationen
PAT	Principal-Agent-Theorie

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Im Evaluationskontext gilt die wissenschaftliche Unabhängigkeit als wesentliche Voraussetzung für Evaluationen (Pleger, Sager, Morris, Meyer, & Stockmann, 2016, S. 1). Gleichzeitig befinden sich Evaluationen automatisch in einem organisationalen und politischen Kontext, indem Auftraggeber von Evaluationen oft ein spezifisches Interesse an den Evaluationsresultaten verfolgen (Barnett & Camfield, 2016, S. 528). Damit stehen Evaluierende oft im Spannungsfeld die Unabhängigkeit von Evaluationen aufrechtzuerhalten und gleichzeitig den Erwartungen der Auftraggeber gerecht zu werden (Pleger et al., 2016, S. 12). Für die Gewährleistung von unabhängigen und objektiven Evaluationsprozessen ist es zentral, dass die Auftraggeber keinen Einfluss auf den Evaluationsprozess ausüben (Pleger & Sager, 2018, S. 166). Evaluierende fühlen sich jedoch häufig unter Druck gesetzt, wobei der Auftraggeber von Evaluationen als Stakeholder mit der grössten Einflussnahme im Evaluationsprozess identifiziert wurde (Morris & Clark, 2012; Morris & Cohn, 1993, 1993; Pleger et al., 2016; Stockmann, Meyer, & Schenke, 2011; Turner, 2003). Die Druckausübung kann viele Formen annehmen und die Einhaltung von Evaluationsstandards und Prinzipien gefährden, wobei die für die Unabhängigkeit von Evaluationen wichtigen Faktoren der Objektivität und wissenschaftlichen Integrität gefährdet werden (Pleger et al., 2016, S. 11). Demgegenüber überrascht die Tatsache, dass zur Auftraggebersicht bislang wenig Forschung vorhanden ist, während sich bisherige Forschung vordergründig auf der Perspektive von Evaluierenden und ihrem Umgang mit ethischen Herausforderungen fokussierte (Barnett & Camfield, 2016; Brown & Newman, 1992; Desautels & Jacob, 2012; Jacob & Boisvert, 2010; Morris, 1999; Morris & Cohn, 1993; Morris & Jacobs, 2000; Pope & Vetter, 1992; Stufflebeam, 1994; Turner, 2003). Um diese eingeschränkte Sicht zu erweitern, wird mit der Untersuchung der Auftraggebersicht zum Verständnis des komplexen Kontextes des Evaluationsprozesses beigetragen, die Forschungslücke reduziert und der Forschungsstand durch diese integrative Sicht weiterentwickelt und vervollständigt.

Pleger und Hadorn (2018) untersuchen in ihrer Studie die Auftraggeberseite in der Schweiz und zeigen, dass die befragten Auftraggeber kaum auf ihre Druckausübung seitens Evaluierenden angesprochen wurden, das Verhältnis zwischen Evaluierenden und Auftraggebern jedoch oft konfliktgeprägt ist. Die Schweizer Studie bietet einen deskriptiven Einblick in die Auftraggebersicht von Evaluationen und fungiert als Vorbildstudie für die vorliegende Masterarbeit mit dem Ziel einen umfassenden Überblick über die

Auftraggeberperspektive der Evaluationslandschaft der Vereinigten Staaten zu schaffen. Dabei wird auf der Studie als methodische Basis aufgebaut und die Befunde durch eine methodische sowie inhaltliche Erweiterung mit Fokus auf der Principal-Agent-Theorie (PAT) validiert. Indem auf die Auftraggeber-Perspektive in den USA fokussiert wird, trägt die Studie dazu bei, ein vollständigeres Portrait von Fehldarstellungen in Evaluationen und Unabhängigkeit von Evaluationen zu schaffen. Darüber hinaus liefert der Ländervergleich zwischen der Evaluationslandschaft der USA und der Schweiz weitere Erkenntnisse über spezifische Gemeinsamkeiten und Unterschiede. Mit dem Fokus auf der Auftraggeberseite wird untersucht, wie Auftraggeber die Unabhängigkeit von Evaluationen beurteilen und welche Rolle die Eigenschaften der Auftraggeber und ihre Beziehung zu Evaluierenden in der Beeinflussung des Evaluationsprozesses spielen. Dabei wird untersucht, inwiefern ein konfliktgeprägtes Principal-Agent-Verhältnis mit der Einflussnahme auf den Evaluationsprozess zusammenhängt und wie divergierende Interessen seitens der Auftraggeber ausgeglichen werden. Zudem wird der Zusammenhang der Vertrautheit mit Evaluationsstandards und der Erwartung an die Unabhängigkeit von Evaluationen analysiert. Ergänzend werden die Wichtigkeit und Wahrnehmung der Unabhängigkeit von Evaluationen, sowie die von den Auftraggebern wahrgenommene eigene Einflussstärke beleuchtet. Weiter werden die von den Auftraggebern wahrgenommenen Konfliktgründe und vorgeschlagenen präventiven Massnahmen untersucht, die zur Schaffung eines fruchtbaren Umfelds für aussagekräftige Evaluationen beitragen.
Basierend auf dieser Bestandesanalyse können in einem nächsten Schritt für die Praxis mögliche Handlungsempfehlungen und präventive Massnahmen zur Gewährleistung und Verbesserung der Unabhängigkeit von Evaluationen resp. des oftmals konfliktgeprägten Principal-Agent-Verhältnis abgeleitet werden. Anhand einer Online-Befragung von Auftraggebern in den USA wird somit die folgende Hauptforschungsfrage untersucht: *Wie wird die Unabhängigkeit von Evaluationen von Auftraggebern in den USA beurteilt? Welche Rolle spielen die Eigenschaften der Auftraggeber und ihre Beziehung zu Evaluierenden in der Beeinflussung des Evaluationsprozesses?*

Nach der Einleitung wird im 2. Kapitel in das Thema der ethischen Herausforderung im Evaluationsprozess und der Unabhängigkeit von Evaluationen anhand des aktuellen Forschungsstands eingeführt. Danach werden mögliche Beeinflussungsformen von Auftraggebern mit Fokus auf dem für den Evaluationskontext relevanten Einflussmodell erläu-

tert. Nach diesem Überblick wird die der Studie zugrundeliegende Principal-Agent-Theorie (PAT) vorgestellt. Basierend auf den theoretischen Annahmen werden das Conceptual Framework und die Hypothesen abgeleitet und die Forschungsfragen vorgestellt. Das 3. Kapitel stellt das zugrundeliegende Forschungsdesign vor. Dabei wird zunächst auf das Material und Sample sowie die Fragebogenentwicklung eingegangen und die Operationalisierung der Konstrukte und Skalenentwicklung präsentiert. Abschliessend werden die Datenauswertungsmethoden erörtert. Im 4. Kapitel werden die Resultate der durch die Online-Befragung gesammelten Daten einerseits in Hinblick auf die Überprüfung der Hypothesen und andererseits im Ländervergleich auf systematische Weise beschrieben. Im 5. Kapitel werden die Resultate diskutiert und die Forschungsfragen schrittweise beantwortet. Das 6. Kapitel beinhaltet die zentralen Schlussfolgerungen, wobei die Implikationen und Limitationen beschrieben werden und die Studie mit einem Ausblick abgerundet wird.

2 Theorie und Hypothesen

Das Ziel dieser Studie liegt in der Untersuchung wie die Unabhängigkeit von Evaluationen von Auftraggebern in den Vereinigten Staaten beurteilt wird. Dabei wird untersucht, welche Rolle die Eigenschaften von Auftraggebern und ihre Beziehung zu Evaluierenden in der Beeinflussung des Evaluationsprozesses spielen. Zu Beginn werden in die Themen der ethischen Herausforderung im Evaluationsprozess und der Unabhängigkeit von Evaluationen anhand von aktueller Forschungsliteratur eingeführt. Anschliessend werden mögliche Beeinflussungsformen von Auftraggebern anhand des für den Evaluationskontext relevanten Einflussmodells beschrieben. Die theoretische Basis dieser Studie bildet die Principal-Agent-Theorie (PAT) nach Jensen und Meckling (1976), welche sich vordergründig mit dem Beziehungsverhältnis zwischen einem *Principal* und *Agent* beschäftigt. Zum Schluss des Kapitels wird die Theorie mit Bezug zum Evaluationskontext vorgestellt und die für die Untersuchung relevanten Annahmen diskutiert. Basierend darauf wird das theoretisch abgeleitete Conceptual Framework mit den dazugehörigen Hypothesen präsentiert und die Forschungsfragen übersichtlich dargestellt.

2.1 Ethische Herausforderung im Evaluationsprozess

International zeigt sich ein wachsendes Interesse in der Beziehung zwischen den Anliegen von Praktikern sowie Politikern und der Forschungsevidenz. Dabei werden nationale Bemühungen zunehmend debattiert, die darin bestehen evidenzbasierte Policies in bestimmten politischen Bereichen wie der sozialen Arbeit, Bildung und Gesundheit zu fördern (Nutley, Morton, Jung, & Boaz, 2010, S. 131). Die Förderung evidenzbasierter Ansätze impliziert für politische Entscheidungsträger bei der Entwicklung von Policies jeweils die beste verfügbare Evidenz zu suchen und dabei besonderen Wert auf nachweisbare Resultate zu legen (Ebd., 2010, S. 132–133). Diese Integration von Evidenz in den politischen Entscheidungsprozess wird unter dem Begriff des *Evidence-based Policy-Making* (EBP) zusammengefasst, wobei EBP sowohl in der Politik als auch in der Forschung zunehmend Beachtung geschenkt wird (Head, 2008; Pleger & Sager, 2018; Sanderson, 2000, 2002; van der Knaap, 2004; Young, Ashby, Boaz, & Grayson, 2002). Die weitgefasste Definition beschreibt den Kern des evidenzbasierten Policy-Makings passend, indem es Entscheidungsträger dabei unterstützt "to make well-informed decisions about policies, programs, projects and practices by putting the best available evidence at the heart of policy development and implementation" (Nutley et al., 2010, S. 133). Die evidenzbasierte Bewegung in der modernen Politik konstituiert sich somit in

der Suche nach relevantem und gleichzeitig nützlichem Wissen damit Probleme identifiziert und gelöst werden können (Head, 2008, S. 1–2). Demokratische Regierungen sind mit zunehmender Infragestellung und Kontrolle öffentlicher Interventionen konfrontiert, sodass ihre Legitimität nicht mehr alleine durch demokratiepolitische Prozesse garantiert ist. Für die Legitimierung ihrer Aktivitäten spielt die Evidenz ihrer Leistungsfähigkeit eine zentrale Rolle (Sanderson, 2002, S. 2). Dabei kommt dieser Evidenz eine doppeldeutige Rolle zu, wobei erstere darin besteht die Verantwortlichkeit der Regierung in Bezug auf die Resultate zu stärken resp. zu zeigen, dass die Regierung effektiv arbeitet. Letztere verfolgt das Ziel der verbesserten Entscheidungsfindung und der Förderung von effektiven Programmen (Ebd., 2002, S. 3). Mit der zunehmenden Wichtigkeit von Evidenz für Regierungen geht auch die Gefahr einher, dass Forschungsergebnisse selektiv verwendet werden, um gerade diejenigen Politiken zu legitimieren, die den individuellen politischen Prioritäten entsprechen (Ebd., 2000, S. 434).

Die aus der EBP resultierende Nützlichkeit von wissenschaftlichen Evaluationsresultaten wird somit für das politische System abgeleitet, wobei die in den Evaluationsresultaten inhärente wissenschaftliche Evidenz zur politischen Vertrauenswürdigkeit beiträgt. Gleichzeitig wird dadurch die politische Entscheidungsfindung erleichtert. In Zusammenhang mit EBP wird angenommen, dass die Evaluationsresultate auf neutraler Evidenz basieren und politisch nicht verzerrt sind (Pleger & Sager, 2018, S. 166). Zentraler Bestandteil von EBP sind Evaluationen, die als "analytical inquiry based on collecting and analyzing evidence, and drawing conclusions and recommendations from this evidence"(Valovirta, 2002, S. 60) beschrieben werden können. Das Prinzip der Generierung objektiver Evidenz durch Evaluationen lässt sich somit eng mit dem Prinzip des EBP in der Politik verknüpfen (Pleger & Hadorn, 2018, S. 2). Bei einer Evaluation werden Evaluierende von Auftraggebern beauftragt, wobei die kollaborative Auftragsbeziehung wissenschaftlichen Prinzipien genügen muss, um evidenzbasierte Resultate zu generieren. Für die Gewährleistung von unabhängigen und objektiven Evaluationsprozessen resp. Evaluationsresultaten ist es zentral, dass die beauftragende Partei keinen Einfluss auf den Evaluationsprozess ausübt (Pleger & Sager, 2018, S. 166). Die Wurzel des Evaluationsbegriffs besteht im Wort „value" und deutet darauf hin, dass Wertungen inhärenter Bestandteil von Evaluationen sind und Evaluationen nicht komplett wertfrei sind (Fox, Grimm & Caldeira, 2017, S. 7; Scriven, 1993 zitiert in Stufflebeam & Coryn, 2014, S. 8). Evaluationen sollen nicht nur inhaltlich auf gewisse Werte wie Effektivität, Effizienz, Kosten, Sicherheit und Nützlichkeit hinweisen, sondern sich auf ein vertretbares Werteset

stützen. Beispiele von professionell definierten Prinzipien sind die „Guiding Principles for Evaluators“ der American Evaluation Association (AEA) oder die „Program Evaluation Standards“ des Joint Committee on Standards (Stufflebeam & Coryn, 2014, S. 8). Die fünf Guiding Principles der AEA beziehen sich dabei auf ethische Verhaltensweisen von Evaluierenden und adressieren „systematic inquiry, competence, integrity, respect for people, and common good and equity“, wobei sich die Evaluationsstandards stärker mit der Evaluationsqualität auseinandersetzen (American Evaluation Association [AEA], 2011, S. 2). Gemäss letzterer sollten Evaluierende die Werte der Standards des Nutzens, der Machbarkeit, Korrektheit, Genauigkeit und der Verantwortlichkeit der Evaluation einhalten (Stufflebeam & Coryn, 2014, S. 8). Die Bedürfnisse der Auftraggeber von Evaluationen werden dadurch befriedigt, dass die durch die Evaluation erlangten Informationen ihre Werturteile stützen. Dabei ist es zentral, dass jenen Informationen diese ganze Bandbreite an Werten zugrunde liegt (Ebd., 2014, S. 14). Evaluationen können nämlich eine erhebliche Wirkung nach sich ziehen. Einerseits können Evaluationen mit unangemessenen Erwartungen aufseiten von involvierten Stakeholdern verbunden sein, andererseits können sie zu unverhofften Evaluationsresultaten führen, was wiederum negative Konsequenzen implizieren könnte. Demzufolge müssen Evaluationen möglichst transparent, glaubwürdig, unparteiisch und legitim sein (Desautels & Jacob, 2012, S. 438).

Auch moderne Verwaltungen beschäftigen sich immer mehr mit ethikrelevanten Fragestellungen in Zusammenhang mit evaluativen Prozeduren. Damit einher gehen stärkere Bestrebungen von professionellen Institutionen, wie der AEA oder der Canadian Evaluation Society, Richtlinien und Standards und damit die Evaluationspraxis zu fördern (Ebd., 2012, S. 437). Aufgrund dieser Bedeutung etablierte sich in der Evaluationsliteratur ein klarer Fokus in der Untersuchung von ethischen Herausforderungen, denen sich Evaluierende stellen müssen und die gleichzeitig die Evaluationsqualität gefährden (Barnett & Camfield, 2016; Brown & Newman, 1992; Desautels & Jacob, 2012; Jacob & Boisvert, 2010; Morris, 1999; Morris & Cohn, 1993; Morris & Jacobs, 2000; Pope & Vetter, 1992; Stufflebeam, 1994; Turner, 2003). Unterschiedliche Studien legen nahe, dass der Evaluationsprozess mit ethischen Herausforderungen verbunden ist. Eastmond (1998) und Scheirer (1998) behaupten, dass die Handlungen von Evaluierenden während dem Evaluationsprozess als Hauptquelle von ethischen Problemen fungieren. Jedoch variiert die wahrgenommene Wichtigkeit und Salienz von ethischen Bedenken beträchtlich unter Evaluierenden (Morris, 1999, S. 17). Morris (2008, S. 2) beschreibt ein ethisches

Dilemma als Situation, bei welcher der Evaluierende bei Fragen der moralischen Verantwortlichkeit eine ethische Wahl zwischen „‘doing the right (good) thing’ or ‘doing the wrong (bad) thing‘“ hat. Trotz dem gestiegenen Bewusstsein, dass ethische Probleme angegangen werden sollen, existiert nur wenig systematische, strukturierte Forschung in Bezug auf ethische Dilemmata im Evaluationsprozess (Desautels & Jacob, 2012, S. 438; Morris, 1999, S. 15). Dank zwei Studientypen konnte das Wissen bezüglich ethischer Verhaltensweisen von Evaluierenden, v.a. in den Vereinigten Staaten, erweitert werden (Desautels & Jacob, 2012, S. 438). Ersterer fokussiert primär darauf wie sich Evaluierende generell verhalten, wenn sie sich mit einer ethischen Frage konfrontiert sehen (Brown & Newman, 1992). Der zweite Projekttyp ergänzt den ersten und untersucht die persönliche ethische Sensitivität der Evaluierenden, indem ihre wahrgenommene Rolle im Evaluationsprozess oder ihre Perspektiven in Bezug auf ethische Fragen näher beleuchtet werden (Morris & Cohn, 1993; Morris & Jacobs, 2000). Brown und Newmanm (1992, S. 655) untersuchen in ihrer Studie wie sich aus Sicht von Evaluierenden und Auftraggebern ethische Prinzipien mit einer Auswahl der durch die Joint Committee on Standards (1981) entwickelten Evaluationsstandards decken. Evaluationsexperten haben dabei häufiger eine Übereinstimmung zwischen den vorgegebenen ethischen Prinzipien und den Standards erreicht als Auftraggeber von Evaluationen und Personen mit keiner oder mässiger Evaluationskenntnis. Die Resultate deuten darauf hin, dass Berufserfahrung und Bildung zu einem gemeinsamen Evaluationsverständnis führen können und dieses gemeinsame Verständnis von entsprechender Einigkeit bezüglich erwarteter Verhaltensweisen zeugt (Ebd., 1992, S. 661). Die Studie von Brown und Newman (1992) lässt keine Schlüsse auf die persönlichen Erfahrungen von Evaluierenden bezüglich ethischer Herausforderungen zu. In der Untersuchung von ethischen Konflikten von randomisiert ausgewählten Mitgliedern der American Psychological Association wurden Respondenten direkt anhand einer offenen Frage nach ethischen Bedenken befragt (Pope & Vetter, 1992, S. 398). Diese Vorgehensweise ermöglicht somit ein genaueres Abbild der Erfahrungen von Evaluierenden mit ethischen Herausforderungen zu generieren, als dies andere Ansätze erlauben (Morris & Cohn, 1993, S. 623). Morris und Cohn (1993, S. 621) nutzen diese respondent-getriebene Methodik im Rahmen einer Befragung von einem randomisierten Mitgliedersample der AEA, um zu untersuchen welchen ethischen Herausforderungen Evaluierende während ihrer Arbeit begegnen. Die Darstellung von Evaluationsresultaten wurde von den Befragten am häufigsten als ethischer Hauptkonflikt beschrieben. In Bezug auf diese generelle Kategorie gaben die Befragten an, dass sie sich durch

Stakeholder – normalerweise den Auftraggeber – unter Druck fühlten gewisse Fakten zu verzerren (Ebd., 1993, S. 628–630). Bezüglich der Fehlinterpretation und dem Missbrauch von Resultaten als Herausforderung gaben mehr als ein Drittel der Befragten an, dass ihr finaler Report unterdrückt oder nicht gebraucht wurde, weiter wurde erwähnt, dass die Evaluationsresultate absichtlich vom Auftraggeber modifiziert oder fehlinterpretiert wurden (Ebd., 1993, S. 631). Für jede Phase des Evaluationsprozesses wurde festgestellt, dass ethische Probleme auftreten oder potentiell auftreten können. Gemäss den untersuchten Wahrnehmungen der Evaluierenden sind den identifizierten ethischen Problemen gemeinsam, dass sich Evaluierende häufig unter Druck gesetzt füllen ihre Rolle als 'Wissenschaftler' zu schwächen. Genauer charakterisiert sich diese Druckausübung darin, dass der Auftraggeber versucht „die grundlegende Mission der wissenschaftlichen Forschung zu unterdrücken, die darin besteht, die Wahrheit zu suchen und zu vermitteln" (Morris & Cohn, 1993, S. 639). Auch bei Turner (2003, S. 1) zeigen die Resultate der Online-Befragung von Mitgliedern der Australasian Evaluation Society, dass ethische Herausforderungen in Zusammenhang mit den Auftraggebern von Evaluationen dominieren. Respondenten spezifizierten diese Herausforderungen durch die Situation, dass Auftraggeber versuchen Evaluationsresultate zu kontrollieren oder zu beeinflussen. Zusätzlich wurde die Druckausübung aufseiten des Auftraggebers genannt, mittels welcher versucht wird gewisse Evaluationsresultate besser oder gar schlechter darzustellen.
Im nachfolgenden Abschnitt wird anhand von aktueller Forschungsliteratur vertieft auf die Thematik der Druckausübung im Evaluationsprozess aufseiten von Auftraggebern eingegangen, die mitunter die Unabhängigkeit von Evaluationen und die reine Evidenz als Voraussetzungen für EBP gefährdet (Pleger & Sager, 2018, S. 168).

2.2 Unabhängigkeit von Evaluationen in Gefahr

Nachdem sich die Evaluationsforschung vorwiegend auf Evaluationsstandards im Allgemeinen und ethische Herausforderungen in der Evaluationspraxis konzentrierte, gelang die Unabhängigkeit von Evaluationen vermehrt in den Forschungsfokus (Pleger et al., 2016, S. 3). Im Evaluationskontext gilt die wissenschaftliche Unabhängigkeit als wesentliche Voraussetzung für Evaluationen, wobei diese mit anderen Voraussetzungen zusammen wiederum durch die Evaluationsstandards und -prinzipien der länderspezifischen Evaluationsgesellschaft definiert werden (Ebd., 2016, S. 1). Das normative Prinzip der Integrität der AEA beschreibt die Bewahrung der Unabhängigkeit im Rahmen von professionellen, ethischen Verhaltensweisen treffend, indem es normativ unterstreicht, dass sich Evaluierende transparent und ehrlich verhalten, um die Integrität der Evaluation zu

gewährleisten (AEA, 2011, S. 3). Die Unabhängigkeit findet sich auch als Grundpfeiler in den „Quality Standards for Development Evaluation“ der OECD (2010) wieder, um die Qualität für den Evaluationsprozess und das Evaluationsprodukt sicherzustellen. Dabei wird die Unabhängigkeit in Zusammenhang mit den übergreifenden Überlegungen und in Bezug auf die Implementation und das Reporting aufgegriffen. Übergreifend wird festgehalten, dass der Evaluationsprozess unabhängig und transparent vom Programmmanagement und Policy-Making ist, um die Glaubwürdigkeit zu erhöhen (Ebd., 2010, S. 6). Im Bereich der Implementation und dem Reporting wird ausgeführt, dass sowohl die verwendete Methodologie als auch aufgetretene Einschränkungen und deren Auswirkungen auf die Evaluation und deren Unparteilichkeit und Unabhängigkeit im Evaluationsreport beschrieben und erklärt werden (Ebd., 2010, S. 13). Der Standard zur Unabhängigkeit von Evaluierenden gegenüber Stakeholdern wird wie folgt definiert (Ebd., 2010, S. 11):

> Evaluators are independent from the development intervention, including its policy, operations and management functions, as well as intended beneficiaries. Possible conflicts of interest are addressed openly and honestly. The evaluation team is able to work freely and without interference. It is assured of co-operation and access to all relevant information.

Für die vorliegende Masterarbeit fungiert dieser Standard als definitorische Basis zur Erklärung, was unter der Unabhängigkeit von Evaluationen verstanden wird. Diese Basis ist insofern relevant, da sie unterschiedliche für die Studie relevante Dimensionen abdeckt, die mit der PAT (siehe Kapitel 2.4) in Verbindung stehen. Genauer unterstreicht sie die Unabhängigkeit der Evaluation von beabsichtigten Begünstigten, in diesem Fall von Auftraggebern von Evaluationen als relevanten Stakeholder. Zudem wird die grundsätzliche Annahme der PAT aufgegriffen, die im Interessenkonflikt zwischen dem Auftraggeber als *Principal* und Evaluierenden als *Agent* besteht. Weiter steht die freie Arbeitsweise des Evaluierenden ohne Intervention resp. Beeinflussungsversuchen aufseiten des Auftraggebers im Vordergrund, die durch Kooperation und Bereitstellung aller relevanten Informationen unterstützt wird. Dazu hält Widmer (2012, S. 130) treffend fest, dass sich die Unabhängigkeit von Evaluationen dadurch charakterisiert, dass „ein Akteur ohne Interferenzen von Dritten frei agieren kann [und damit] eine Unterscheidung in *Prinzipal* und *Agent* hinfällig wird“. Der Begriff der Unabhängigkeit geht damit von einer Negation resp. dem Fehlen einer Abhängigkeit aus und widerspricht den grundsätzlichen

Annahmen der PAT. Eine gegenseitige Beeinflussung ist jedoch inhärenter Bestandteil jeder sozialen Interaktion, was impliziert, dass nicht dichotom, sondern graduell zwischen Abhängigkeit und Unabhängigkeit unterschieden werden kann. Eine eindeutige Unterscheidung ist problembehaftet, da es nicht zentral ist, ob eine Abhängigkeit zwischen den Akteuren besteht, sondern viel mehr wie stark diese ausgeprägt ist (Ebd., 2012, S. 130). Das Verhältnis zwischen dem Auftraggeber (*Principal*) und Evaluierenden (*Agent*) als abhängigen Akteur ist von einem bestimmten Stärkegrad an Abhängigkeit resp. einer Einflussnahme aufseiten des Auftraggebers geprägt. Diese Einflussnahme oder Druckausübung geht über eine lediglich soziale Interaktion hinaus, woraus abgeleitet angenommen wird, dass die Unabhängigkeit im Evaluationskontext tangiert oder gefährdet wird. Das beschriebene Abhängigkeitsverhältnis ist nicht nur unidirektional von einer Dependenz geprägt, sondern findet wechselseitig statt und charakterisiert sich somit durch eine gewisse Interdependenz beider Akteure (Ebd., 2012, S. 131). Der Druck auf Evaluierende gilt als wichtiger Typ von ethischer Herausforderung während dem Evaluationsprozess (Morris, 1999, 2007), wobei verstärkte Aufmerksamkeit darauf gerichtet wurde, was passiert, wenn Stakeholder unabhängige Evaluationsresultaten durch ihre Druckausübung beeinflussen (Pleger et al., 2016). Wie bereits erwähnt kommt dem Auftraggeber oftmals die Hauptrolle in der Beeinflussung des Evaluationsprozesses zu. Diese konstituiert sich aus der Macht und Position heraus, welche den Auftraggeber damit befähigt einzelne Aspekte im Evaluationsprozess wie bspw. den Umfang, die Methodologie und Ressourcen, aber auch einzelne Ergebnisse zu beeinflussen oder gar zu bestimmen (Barnett & Camfield, 2016, S. 532). Evaluierende stehen damit im Spannungsfeld den Erwartungen der Auftraggeber gerecht zu werden und gleichzeitig die Unabhängigkeit von Evaluationen aufrechtzuerhalten. Die Unabhängigkeit von Evaluationen ist in Gefahr, sobald Auftraggeber Druck auf die Evaluierenden ausüben, um z. B. gewisse Evaluationsresultate zu verändern oder fehlzuinterpretieren (Pleger et al., 2016, S. 12). Dieses Spannungsfeld liegt gewissermassen in der Natur von Evaluationen. Im Vergleich zur sozialwissenschaftlichen Forschung teilen sie sich zwar die gleiche Methodologie, grenzen sich in Bezug auf die Interessen ihrer Stakeholder zur Forschung ab. Evaluationen spielen automatisch in einem organisationalen und politischen Kontext, wobei sich Stakeholder von Evaluationen meistens ein spezifisches Interesse an den Evaluationsresultaten verfolgen (Barnett & Camfield, 2016, S. 528). Genauer werden Evaluationen nicht automatisch in Gang gesetzt, sondern normalerweise direkt durch dritte Parteien wie Beamte oder Politiker mit einem gewissen Interesse innerhalb des politischen Kontextes beauftragt (Pleger

& Hadorn, 2018, S. 2). Demzufolge ist der Evaluationsprozess trotz seiner wissenschaftlichen Verankerung dennoch geprägt von Verzerrungen und kann niemals vollständig neutral sein (Desautels & Jacob, 2012, S. 437; Eliadis, Furubo & Jacob, 2010). Barnett und Camfield (2016, S. 529) argumentieren, dass sich die politische Ökonomie verzerrend auf unterschiedliche Aspekte des Evaluationsprozesses auswirken und somit bereits die Formulierung von Evaluationsfragen oder Entscheidungen in Bezug auf Ressourcen- und Methodenwahl tangieren kann. Im Vergleich zur Forschung wird zudem dargelegt, dass Evaluationen einen grösseren und direkteren Praxisnutzen für Entscheidungsträger haben und die involvierten Stakeholder – insb. die Auftraggeber – während dem Evaluationsprozess stärker beteiligt sind.

Die komparative empirische Analyse von vier Studien aus den Vereinigten Staaten, Grossbritannien, Deutschland und der Schweiz (Morris & Clark, 2012; Pleger & Sager, 2016b; Stockmann et al., 2011; THE LSE GV314 GROUP, 2013) stellt länderspezifische Befragungsresultate bezüglich der Druckausübung von Auftraggebern auf Evaluierende in vergleichenden Zusammenhang, wobei zwei Befunde übereinstimmend dominieren. Erstens wurde der Auftraggeber von Evaluationen als Stakeholder mit der grössten Einflussnahme im Evaluationsprozess identifiziert. Zweitens zeigte sich, dass viele Respondenten Evaluationen nicht als unabhängig wahrnehmen (Pleger et al., 2016, S. 1). Daraus kann abgeleitet werden, dass Auftraggeber versuchen ihre eigenen Interessen und Präferenzen in den Evaluationsprozess einzubringen, was wiederum verzerrte Evaluationsresultate zur Folge hat, die mitunter dazu führen, dass Evaluationen als nicht unabhängig wahrgenommen werden. Die verzerrten Evaluationsresultate dienen dem Auftraggeber schliesslich zur Entscheidungsfindung, da sie nicht zuletzt mit den eigenen Präferenzen korrespondieren. Die Druckausübung kann viele Formen annehmen und die Einhaltung von Evaluationsstandards und Prinzipien gefährden. Genauer kann sich diese Intervention auch hier in der Aufforderung zur Änderung von einzelnen Satzteilen bis zur Verzerrung von Evaluationsresultaten äussern, wodurch die für die Unabhängigkeit von Evaluationen wichtigen Faktoren der Objektivität und wissenschaftlichen Integrität gefährdet werden (Ebd., 2016, S. 11). Obwohl die Perspektive der Evaluierenden anhand der Studie übersichtlich dargestellt wurde, können dennoch keine Aussagen über die Wahrnehmung der Auftraggeber hinsichtlich ihrer Einflussnahme gemacht werden (Ebd., 2016, S. 12). Damit lässt sich die Relevanz der vorliegenden Untersuchung begründen, die Auftraggeberseite zu beleuchten und vertieft zu untersuchen, wie die Unabhängigkeit von Evaluationen aus Auftraggebersicht beurteilt wird.

Demgegenüber können Evaluierende auf unangemessene Weise bestrebt sein ihren Stakeholdern diejenigen Resultate zu berichten, die aus ihrer Sicht von den Auftraggebern als erwünscht wahrgenommen werden (Morris, 2007, S. 413). Die sozialpsychologische Forschung mit Fokus auf Attributionsprozesse liefert einen Erklärungsversuch für Individuen aus westlichen Gesellschaften. Diese Forschung deutet darauf hin, dass Individuen dazu tendieren ihr eigenes Verhalten nicht als Problemursache zu sehen, sondern externe Quellen für die Probleme verantwortlich zu machen (Kelley, 1973). Ein solcher bei den Evaluierenden auftretender Attributionsprozess würde entsprechend implizieren, dass Evaluierende selbst diese Art von Fehldarstellungen der Ergebnisse vornehmen. Diese Implikation wird entsprechend vor dem Hintergrund betrachtet, dass den Auftraggebern so oft vorgeworfen wird, dass sie die Evaluierende zu dieser Handlung beeinflussen (Morris, 2007, S. 413). Folglich wird damit die Relevanz für vorliegende Studie umso mehr verstärkt, dass die Rolle der Eigenschaften und des Beziehungsverhältnisses von Evaluierenden und Auftraggebern in der Einflussnahme des Evaluationsprozesses untersucht wird. Entsprechend liegt es im Forschungsinteresse zu untersuchen, ob Auftraggeber tatsächlich zu Fehldarstellungen ermuntern oder inwiefern ihre Beeinflussungsform eingeordnet werden kann. Morris (2007) betont ausserdem, dass es für Fehldarstellungen auch andere Gründe als lediglich Druckausübung gibt und sich die Forschung sowohl auf die Evaluierenden- als auch die Perspektive der Auftraggeber fokussieren soll. Die Tatsache, dass sich bisher wenig Forschung auf die Auftraggeberperspektive fokussierte, mag einerseits darum erstaunen, da der Unabhängigkeit von Evaluationen generell zunehmend Aufmerksamkeit geschenkt wurde. Andererseits liegt die erwähnte, konsistente Erkenntnis vor, dass Auftraggeber eine Schlüsselrolle in der Druckausübung auf Evaluierende einnehmen und damit die Unabhängigkeit von Evaluationen gefährden, wobei die verzerrte Evidenz wiederum die grundlegende Idee des EBP untergräbt (Pleger & Hadorn, 2018, S. 3). Die explorative Pilotstudie von Pleger und Hadorn (2018) setzt an dieser Problematik an und versucht die direkt damit verbundene Hauptforschungsfrage zu beantworten, wie Auftraggeber ihre Beziehung zu Evaluierenden wahrnehmen. Genauer werden die Fragen untersucht, ob sich die Auftraggeber ihrer Druckausübung tatsächlich bewusst sind und wie sie die Wichtigkeit von unabhängigen Evaluationen einschätzen. Darüber hinaus wird untersucht, welche Massnahmen Auftraggeber zur Schaffung eines fruchtbareren Evaluationsumfelds vorschlagen. Die Untersuchung der Auftraggeberseite ist unverzichtbar, um einerseits die Auftraggeber und ihre Sichtweisen bes-

ser zu verstehen, was gegebenenfalls den Dialog zwischen Evaluierenden und Auftraggebenden bereichern kann (Pleger & Hadorn, 2018, S. 2). Anhand einer Online-Befragung liefert die Pilotstudie einen ersten deskriptiven Einblick in die Auftraggeberperspektive und vergleicht die Wahrnehmungen von Schweizer Evaluierenden in Hinblick auf die Unabhängigkeit von Evaluationen basierend auf den Ergebnissen von Pleger und Sager (2016b) mit denjenigen von Auftraggebern in der Schweiz. Gemäss der Studie wurden die befragten Auftraggeber kaum auf ihre Druckausübung seitens der Evaluierenden angesprochen, obwohl das Verhältnis zwischen Evaluierenden und Auftraggebern oft konfliktgeprägt ist. Als häufiger Konfliktgrund wurde der Mangel am gegenseitigen Verständnis zwischen den beiden Parteien identifiziert. Aufgrund dieser Wichtigkeit und der bestehenden Forschungslücke fokussiert sich die Masterarbeit ebenso auf der Untersuchung der Unabhängigkeit von Evaluationen als ethische Herausforderung aus Sicht von Auftraggebern. Dabei fungiert die Schweizer Studie als Vorbildstudie für die vorliegende Untersuchung mit dem Ziel einen umfassenden Überblick der Auftraggeberperspektive in der US-Evaluationslandschaft zu schaffen. Dabei wird auf der Studie als methodische Basis aufgebaut und die Befunde durch eine methodische und inhaltliche Erweiterung mit Fokus auf der PAT validiert, um weitere, länderübergreifende Erkenntnisse für die US-Evaluationslandschaft zu erlangen. Zudem werden die Studienresultate beider Länder in Bezug auf zentrale Aspekte verglichen, was zusätzliche Aussagen über Gemeinsamkeiten und Unterschiede der Auftraggeber über die Ländergrenzen hinweg ermöglicht. Somit liefert die Studie erstmalige Ergebnisse zur Auftraggeberperspektive in den USA und trägt damit zur Weiterentwicklung dieser Forschungslücke entscheidend bei. Um die Einflussnahme auf den Evaluationsprozess von Auftraggebern besser zu verstehen, werden im nächsten Abschnitt mögliche Beeinflussungsformen von Auftraggebern erörtert.

2.3 Beeinflussungsformen von Auftraggebern

Bislang wurde einseitig auf eine negative Einflussnahme der Auftraggeber mit negativen Konsequenzen für die Unabhängigkeit von Evaluationen eingegangen. Die wissenschaftliche Unabhängigkeit einer Evaluation wird aber durch die Einflussnahme eines Auftraggebers nicht per se reduziert (Pleger & Sager, 2016a, S. 46). Die Einflussnahme eines Auftraggebers, bspw. eines Politikers, kann nicht als negativ angenommen werden. Das wechselseitige Zusammenspiel zwischen dem politischen Prozess – oder dem Evaluationsprozess – und den Evaluationsresultaten kann sogar als sich gegenseitig begünstigende Beziehung angesehen werden. Dieses Verhältnis kann bei einer Einflussnahme, die aus einem Wissensaustausch besteht, zu einer Weiterentwicklung oder gar Verbesserung

führen (Pleger & Sager, 2018, S. 168). Um die Evaluationsqualität zu verbessern, existiert eine grosse Bandbreite an positiven Einflussformen während des Evaluationsprozesses. Die dadurch steigende Qualität der Evaluation kann wiederum dazu führen, dass die Legitimität der (politischen) Entscheidung, die aus ebendieser Evaluation hervorgebracht wurde, gestärkt wird (Pleger & Sager, 2016a).

Perrin (2018) diskutiert mit Referenz auf seine eigenen Erfahrungen, unter welchen Bedingungen und in welcher Form Rückmeldungen und Änderungsvorschläge aufseiten von Stakeholdern angebracht sind. Die normative Behauptung, dass Evaluierende immun gegenüber Kritik und Feedback sein sollen, ordnet der Autor als unvernünftig ein, wobei dies jedoch der Realität entspreche. Jeder Autor unabhängig von der Publikationsart unterliegt einer Art Review, wobei die damit verbundenen Anpassungsvorschläge und Kommentare konstruktiver Natur sein können. Konstruktive Beispiele dafür können Änderungsvorschläge in Bezug auf Resultate, Schlussfolgerungen und Empfehlungen sein, wenn diese nicht als logisch erscheinen oder sogar gewisse Schlussfolgerungen oder Empfehlungen im Report fehlen, die aus den Evaluationsresultaten resultieren. Zudem sind Evaluierende nicht unfehlbar und arbeiten nicht alle nach den gleich hohen Evaluationsstandards. Ausserdem können den kompetentesten Evaluierenden Fehler unterlaufen. Wird die Anzahl möglicher Evaluationsdesigns und -ansätze berücksichtigt, erstaunt es nicht, dass unterschiedliche Resultate und Schlussfolgerungen daraus resultieren können. Demzufolge kann es angebracht sein Fragen in Bezug auf die methodische Herangehensweise und die damit verbundenen Ergebnissen innerhalb des Reports zu stellen (Ebd., 2018, S. 2). Evaluierende sind mit der Herausforderung konfrontiert, dass sie Evaluationsresultate generieren müssen, die sowohl sinnvoll für die Stakeholder sind und zugleich zu einem Nutzen führen oder diesen zumindest begünstigen. Entsprechend schwierig ist die Beantwortung der Frage, welche Änderungsvorschläge aufseiten der Auftraggeber angemessen sind und welche nicht (Ebd., 2018, S. 4). Zur Veranschaulichung legitim betrachteter Typen von Änderungsvorschlägen nennt Perrin einige Beispiele. Darunter fällt die Korrektur von faktischen Fehlern, redaktionelle Änderungen, eine der Zielgruppe angepasste Sprache, Änderungen emotionaler Sätze, welche entgegen der eigenen Intention interpretiert werden, ein informativer Fokus auf Prioritäten und die Minimierung von der Evaluation resultierenden, nicht intendierter Konsequenzen (Ebd., 2018, S. 4–6). Um Ordnung in die vielfältigen negativen, aber auch positiven Einflussformen der Auftraggeber zu bringen, wird im nächsten Abschnitt das für den Evaluationskontext relevante Einflussmodell vorgestellt.

2.3.1 Das Einflussmodell im Kontext von Evaluationen (BUSD)

Im Rahmen der Studie von Pleger und Sager (2018) wird ein heuristisches Einflussmodell im Kontext von Evaluationen entwickelt, das neben der negativen Beeinflussung auf Evaluierende, erstmals auch die positive Beeinflussung dieser untersucht. Das heuristische sog. BUSD-Modell stützt sich auf die Annahmen des interaktiven EBP-Modells. Dabei wird angenommen, dass sich die Forschung und das Policy-Making gegenseitig beeinflussen und es zu einem Zusammenspiel zwischen Entscheidungsträger und Forschenden kommt. Dabei beeinflussen beide Akteure die Agenda des anderen (Pleger & Sager, 2018, S. 168; Young et al., 2002). Die durch die Forschung generierte Evidenz wird jeweils der politischen Entscheidungsfindung vorgelagert und kann nur gewährleistet werden, wenn wissenschaftliche Voraussetzungen bei der Durchführung von Evaluationen berücksichtigt werden. Demnach kann die wissenschaftliche Qualität durch die Politik und die Evaluationsbeziehung beeinflusst werden (Pleger & Sager, 2018, S. 168).

Die Spannweite der Einflussversuche seitens des Auftraggebers wird wie bereits erwähnt gemäss unterschiedlichen Studien als gross bewertet (Morris & Clark, 2012; Pleger & Sager, 2016b, 2016a; Stockmann et al., 2011). Zugleich weisen Pleger und Sager (2018, S. 167) darauf hin, dass die Resultate mit Sorgfalt zu interpretieren sind, da sie aus unterschiedlichen Verständnissen vom Begriff *Einfluss* oder *Druck* seitens der Evaluierenden resultieren. Um diesbezüglich Klarheit zu schaffen, behandeln die Autoren des BUSD-Modells die beiden Konzepte der ethischen Herausforderung als funktionale Äquivalente, obwohl beide Terme eine unterschiedliche Perspektive beleuchten. *Druck* wird dabei als Handlung seitens der beeinflussenden Partei auf den Evaluierenden definiert, wobei sich *Einfluss* ebenso auf den Evaluierenden bezieht, sich aber als resultierenden Effekt aufgrund des *Drucks* ergibt (Ebd., 2018, S. 167). Wie die erwähnte Studie vorschlägt, werden auch in vorliegender Arbeit die beiden Terme *Einfluss* und *Druck* resp. *Einflussnahme* und *Druckausübung* als funktionale Äquivalente behandelt und mit dem Überbegriff der ethischen Herausforderung während dem Evaluationsprozess zusammengefasst. Da Einflussversuche nicht nur negativer (Morris & Clark, 2012; Pleger et al., 2016; Stockmann et al., 2011), sondern wie die Studien von Pleger und Sager (Pleger & Sager, 2016b, 2016a) sowie Perrin (2018) zeigen, auch positiver Natur sein kann, eignet sich die Anwendung des von den Autoren entwickelten BUSD-Modells. Das Modell berücksichtigt diese ambivalenten Eigenschaften mit und lässt eine klare Unterscheidung zwischen positivem und negativem Einfluss zu (Pleger & Sager, 2018, S. 167). Diese Unterscheidung

scheint unerheblich für das demokratische Outcome von Evaluationen, da daraus Handlungsempfehlungen zur Prävention von negativem Einfluss resp. zur Förderung von positivem Einfluss abgeleitet werden können (Ebd., 2018, S. 168).

Unabhängigkeit von Evaluationen kann aus zweierlei Winkel betrachtet werden: Einerseits aus der formalen Unabhängigkeit, die sich auf die strukturelle Kontrollfreiheit bezieht, andererseits aus der substanziellen Unabhängigkeit, auf welche das Modell Bezug nimmt. Die substanzielle Unabhängigkeit charakterisiert sich durch die Objektivität, die bei der wissenschaftlichen Arbeitsweise notwendig ist und von Verzerrungen und Einflüssen auf die Evaluationsvorgänge und -resultate befreit ist (Pleger & Sager, 2016c, S. 8). Ein Vorteil des BUSD-Modells manifestiert sich darin, dass es auf die Principal-Agent-Situation während dem gesamten Evaluationsprozess angewendet werden kann, da Druck und Einfluss auf jeder Ebene innerhalb des Evaluationsprozesses auftreten können. Aufgrund der möglichen Übertragbarkeit der Studienresultate auf andere Länder mit einer ausgeprägten Evaluationskultur, fungiert das Modell als relevante Basis für die vorliegende Arbeit (Pleger & Sager, 2018, S. 172). Das BUSD-Modell richtet sich insofern an Evaluierende, dass sie die Valenz des auf sie ausgeübten Einflusses systematisch bewerten und als negativ oder positiv einordnen können. Das BUSD-Modell (siehe Abbildung 1) steht für *Betterment, Undermining, Support* und *Distortion* und fasst somit die vier grundsätzlichen Beeinflussungsformen zusammen, die anhand zweier Dimensionen dargestellt werden.

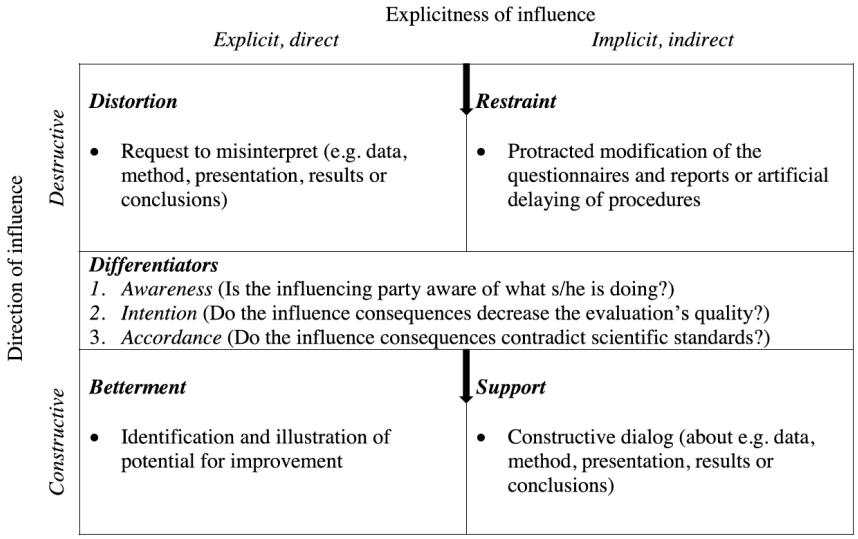

Abbildung 1: BUSD-Modell mit den vier Beeinflussungsformen und Differentiators (Pleger & Sager, 2018)

Die erste Dimension bildet die Beeinflussungsintention (*Direction of influence)* (Pleger & Sager, 2016a) dichotom ab und fängt somit ein, ob die Beeinflussungsform in Bezug

auf die Evaluationsqualität positiver oder negativer Natur ist. Die beiden Pole sind disjunkt, wobei die Beeinflussungsrichtung somit entweder konstruktiv oder destruktiv sein kann. Die Einordnung der Beeinflussungsform in die jeweilige Beeinflussungsrichtung hängt einerseits von der zugrundeliegenden Einflussabsicht und andererseits davon ab, ob die Beeinflussung der evidenzbasierten Forschung nützlich ist (Pleger & Sager, 2018, S. 169). Die zweite Dimension besteht aus dem Entfaltungsgrad von Beeinflussung (*Explicitness of influence)*(Pleger & Sager, 2016a), welche vom Pol des expliziten zum Pol des impliziten Einflusses reicht, jedoch nicht disjunkt ist. Der explizite Einfluss kann als direkte, offensichtliche Druckausübung beschrieben werden und bezieht sich auf die Frage, inwieweit die Unabhängigkeit von Evaluationen beeinflusst wird. Der implizite Einfluss dagegen konstituiert sich vielmehr im indirekten, subtilen Einfluss, wobei die Abgrenzung zum expliziten Einfluss nicht klar, sondern eher graduell gemacht werden kann (Pleger & Sager, 2018, S. 169). Werden die zwei Dimensionen mit ihren vier Ausprägungen kombiniert, können daraus die vier unterschiedlichen Beeinflussungstypen *Betterment, Undermining, Support* und *Distortion* (BUSD) abgeleitet werden. Dabei können die beiden Einflusstypen *Betterment* und *Support* als positive und *Undermining* und *Distortion* als negative Einflussformen eingeordnet werden. Der Beeinflussungstyp *Distortion* gilt als höchster Interventionsgrad in die Unabhängigkeit von Evaluationen und charakterisiert sich durch eine direkte, negative (destruktive) Beeinflussung. Genauer sind darunter direkte Aufforderungen durch Stakeholder gemeint, die jede Art von Fehldarstellung oder Ungenauigkeit mit sich bringen. Damit widerspricht dieser Typ fundamentalen Evaluationsstandards und der wissenschaftlichen Integrität. Letztere stellen eine Gefährdung der wissenschaftlichen Unabhängigkeit dar. Als Beispiele können dabei Aufforderungen sowohl zum Weglassen, zur Veränderung oder Kürzung von gewissen Stellen im Evaluationsbericht, als auch Aufforderungen Methoden, Daten, Resultate, Schlussfolgerungen oder die Präsentation falsch zu interpretieren, genannt werden (Ebd., 2018, S. 170). Der Beeinflussungstyp *Undermining* ist ebenfalls negativer (destruktiver) Natur, charakterisiert sich aber im Vergleich zu *Distortion* durch seine implizite, diskretere Art der Beeinflussung und scheint in der Evaluationspraxis häufig aufzutauchen. Dieser Typus grenzt sich insofern von *Distortion* ab, dass die beeinflussende Partei dem Evaluierenden Vorschläge anstatt Aufforderungen unterbreitet. Bspw. kann dies ein indirekter, verbaler Einfluss sein, der sich in einem Verweis gegenüber den Evaluierenden äussert, wie andere Evaluierende etwas gemacht hätten. Dieser Verweis dient indirekt dazu,

um darauf hinzuweisen, dass etwas bevorzugt anders gemacht worden wäre. Auch mangelhaft formulierte Anforderungsspezifikationen, unanwendbare methodische Richtlinien, das Neuverhandeln von Gebühren oder Zurückhalten wichtiger Informationen kann diesem Typ zugeordnet werden. In Bezug auf den Evaluationsprozess gelten langwierige Anfragen die Berichte oder Fragebögen zu modifizieren, abzuändern oder künstliche Verzögerungen von Prozessen ebenfalls als *Undermining*. Der Beeinflussungstyp *Betterment* dagegen charakterisiert sich in einer direkten, positiven (konstruktiven) Beeinflussung und erhöht die Evaluationsqualität, ohne dabei wissenschaftliche Standards zu missachten oder zu untergraben. Werden Verbesserungspotenziale proaktiv identifiziert und illustriert oder tauschen sich Auftraggeber mit den Evaluierenden über die Präsentation der Ergebnisse aus, um deren Qualität oder das Zielgruppenverständnis zu verbessern, liegt die Beeinflussungsform *Betterment* vor. In Bezug auf das letzte Beispiel ist jedoch zentral, dass die Resultate dabei nicht verzerrt werden (Ebd., 2018, S. 170). Der vierte Beeinflussungstyp *Support* gilt als meist erwünscht und erfasst die Kollaboration zwischen dem Evaluierenden und Auftraggeber, wobei die Evaluationsqualität – ähnlich wie bei *Betterment* – vom Einfluss des Stakeholders profitiert. Letzterer handelt nicht eigennützig, sondern strebt eine Optimierung der Evaluationsqualität an. Dieser Beeinflussungstyp unterscheidet sich vom *Betterment* v.a. in Bezug auf die Einflussintensität des Stakeholders. Als Beispiel dafür fungiert ein konstruktiver Dialog über unterschiedliche Bestandteile der Evaluation wie die Methoden, Daten, Resultate, Schlussfolgerungen oder die Präsentation (Ebd., 2018, S. 170).

Evaluierende sollen jeweils bestrebt sein, wissenschaftliche Richtlinien und Evaluationsstandards aufrecht zu erhalten und negative in positive Einflüsse umzuwandeln. Wird das Modell aus ethischer Perspektive betrachtet, kommt der Unterscheidung zwischen konstruktiven und destruktiven Einfluss eine grössere Bedeutung zu als der Unterscheidung zwischen explizitem, direktem und implizitem, indirektem Einfluss (Ebd., 2018, S. 170). Entsprechend liegt der Forschungsfokus der vorliegenden Arbeit auf ebendieser ersten Dimension. Bevor Evaluierende den positiven in negativen Einfluss transformieren können, zeigt sich in der Evaluationspraxis jedoch das grundsätzliche Problem überhaupt zwischen positivem und negativem Einfluss zu unterscheiden und die Beeinflussung als destruktiv oder eher konstruktiven Dialog einzuschätzen (Pleger & Sager, 2016a, S. 38). In Bezug auf die Klassifizierung des Einflusses wurden drei Ursachen identifiziert, die es Evaluierenden erschweren die Einflussvalenz des Auftraggebers einzuordnen. Die erste Problemursache besteht darin, dass sich Auftraggeber oft nicht ihrem Druck bewusst sind,

den sie auf Evaluierende ausüben. Zweitens ist es für die Evaluierenden schwierig abzuschätzen welche Intention der Auftraggeber mit seiner Druckausübung verfolgt. Drittens kann die Beeinflussung hilfreiche und positive Konsequenzen nach sich ziehen, wobei die Beeinflussung nur dann destruktiver Natur ist, wenn deren Folgen wissenschaftliche Standards wie bspw. Evaluationsstandards verletzt. Um Evaluierende in der Evaluationspraxis bei dieser Unterscheidung zu unterstützen, wurden abgeleitet von den drei Problemursachen sog. *Differentiators* entwickelt (Pleger & Sager, 2018, S. 170). Die drei *Differentiators* (siehe Abbildung 1) mit der Bezeichnung *Awareness, Intention* und *Accordance* unterstützen nicht nur Evaluierende bei der Unterscheidung, ob eine Einflussnahme destruktiv oder konstruktiv ist, sondern tragen gleichzeitig zur theoretischen Forschung im Bereich der Unabhängigkeit von Evaluationen bei. Dabei dienen die hierarchisch geordneten und aufeinander aufbauenden *Differentiators* lediglich zur Bestimmung der Valenz einer Beeinflussung und geben keine Auskunft darüber zu welchem der vier Einflussformen diese eingeordnet wird. Jeder *Differentiators* beinhaltet eine Frage, die den Evaluierenden zur Bestimmung der Einflussvalenz dienen. Je mehr Fragen zugestimmt werden kann, desto destruktiver fällt der Einfluss aus. Bei der Verneinung aller drei Fragen gilt umgekehrt, dass der Einfluss konstruktiver Natur ist. Der Differentiator *Awareness* stellt die Frage, ob sich die beeinflussende Partei ihrer Handlungen bewusst ist, derjenige der *Intention*, ob die Folgen der Beeinflussung die Evaluationsqualität mindern und derjenige der *Accordance*, ob die Beeinflussungsfolgen wissenschaftlichen Standards widersprechen. Der Differentiator *Accordance* wirkt sich zudem am stärksten auf die Dimension der Beeinflussungsintention aus (Ebd., 2018, S. 171).

In vorliegender Arbeit wird das BUSD-Modell insofern weiterentwickelt, dass es erstmals auf die Auftraggeberseite adaptiert wird und zur Eigenbewertung der Einflussvalenz von Auftraggebern verwendet werden kann. Durch die Anwendung des Modells auf die Auftraggeberseite, unterstützt das Modell Auftraggeber dabei ihren eigenen Einfluss auf Evaluierende einzuordnen und zu bewerten. Durch diesen Perspektivenwechsel verändern sich die im Modell inhärenten, impliziten Annahmen geringfügig. Vorher haben Evaluierende den fremden, äusseren Einfluss auf ihre eigene Arbeit bewertet, nun bewerten Auftraggeber ihren eigenen, äusseren Einfluss auf die fremde Arbeit von Evaluierenden. Ob eine Einflussnahme auf die eigene oder fremde Arbeit bewertet wird, stellt einen erheblichen Unterschied dar. Im Gegensatz zur eigenen Arbeit, deren Qualität durchaus einschätzbar ist, ist dies für die fremde Arbeit beinahe unmöglich. Somit stellt sich die

Frage nach der Referenzgrösse in Hinblick auf die fremde Arbeit. Die im Modell inhärenten Annahme implizieren, dass davon ausgegangen wird, dass die „fremde“ Arbeit des Evaluierenden per se den Evaluationsstandards entspricht und als wissenschaftlich integer angenommen wird. Denn nur bei Bekanntheit dieser Referenzgrösse kann das Modell auf die Auftraggeberperspektive angewendet werden. Diese Annahme stellt insofern eine Limitation dar, da sie in der Realität nicht zwingend gegeben sein muss. Die Qualität der Evaluationsarbeit variiert beträchtlich zwischen Evaluierenden, wobei sich jeder als Evaluator ausgeben kann. Dies hängt damit zusammen, dass die Evaluationsdisziplin bislang keine universelle Übereinkunft bezüglich Leitprinzipien, ethischer Richtlinien und notwendigen Kompetenzen gefunden hat (Picciotto, 2019, S. 94). Trotz dieser Limitation trägt das Modell definitiv zur Erweiterung der Evaluationsforschung in Bezug auf die Auftraggeberseite bei und liefert das Instrumentarium zur Untersuchung der Beeinflussungsformen von Auftraggebern. Die Relevanz des BUSD-Modells lässt sich aus dem Zusammenspiel von Evaluationsresultaten und dem politischen Kontext ableiten. Das Modell wird vorliegend nicht nur auf den politischen Kontext, sondern auf das grundsätzliche Umfeld übertragen, wo Evaluationsprozesse in den Vereinigten Staaten stattfinden. Neben dem politischen Kontext werden auch der private und öffentliche Sektor, der NPO-Sektor und der Bereich von Hochschulen und Universitäten beleuchtet. Genauer wird die PAT im Kontext von Evaluationen validiert, wobei der Fokus auf dem Auftraggeber von Evaluationen – dem *Principal* – liegt, der den Evaluierenden (*Agent*) in seiner Evaluationstätigkeit beeinflusst. Das BUSD-Modell fungiert somit als Möglichkeit zur Operationalisierung der Einflussvalenz (siehe Kapitel 3.3) sowie des Entfaltungsgrades der Beeinflussung von Auftraggebern, was mitunter Aussagen über die Beeinflussungsform und somit die Unabhängigkeit von Evaluationen zulässt.

Zusammenfassend liegt der vorliegende Forschungsfokus somit auf der Untersuchung der substanziellen Unabhängigkeit, die durch die Principal-Agent-Situation durch den Auftraggeber (*Principal*) während dem Evaluationsprozess begünstigt resp. gefährdet werden kann. Die Untersuchung bezieht sich somit auf die Principal-Agent-Situation im Evaluationsprozess, ohne dass die ursächliche Motivation für die Evaluation betrachtet wird. Strategische Stakeholderinteressen in einer Evaluation werden nur dann betrachtet, wenn sie in der Beziehung zwischen *Principal* und *Agent* während der Evaluationstätigkeit zustande kommen (Pleger & Sager, 2018, S. 167). Im nachfolgenden Abschnitt wird genauer auf die PAT und ihre Annahmen eingegangen.

2.4 Principal-Agent-Theorie

In folgendem Abschnitt wird die für diese Studie zentrale Principal-Agent-Theorie (PAT) vorgestellt, welche als theoretische Basis die Beziehung zwischen Auftraggebern (*Principal*) und Evaluierenden (*Agent*) im Evaluationsprozess beschreibt. Die Theorie liefert anhand verschiedener Annahmen logische Prognosen darüber wie sich rationale Individuen in der sog. Principal-Agent-Beziehung verhalten (siehe Abbildung 2) (Wright, Mukherji, & Kroll, 2001, S. 414).

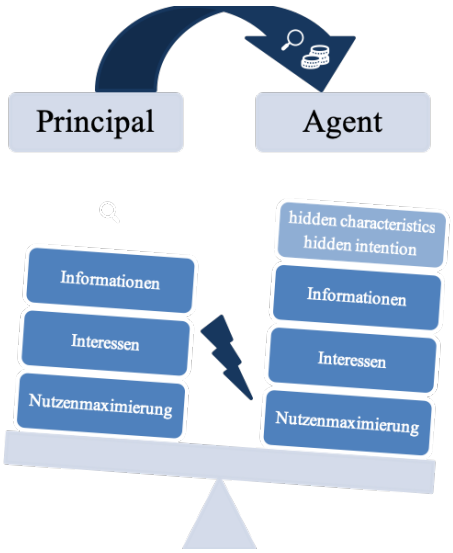

Abbildung 2: Theoretische Annahmen der Principal-Agent-Theorie (eigene Darstellung)

Die soziale Interaktion während dem Evaluationsprozess zwischen dem Auftraggeber – dem *Principal* – und dem Evaluierenden – dem *Agent* – ist geprägt durch ein Austausch, der unvermeidlich eine gegenseitige Beeinflussung impliziert und im Rahmen einer wechselseitigen, bidirektionalen Interdependenz stattfindet. Dieses Abhängigkeitsverhältnis zeichnet sich dadurch aus, dass der *Agent* als abhängiger Akteur in seinem Handlungsspielraum eingeschränkt wird. Aufgrund dieser Abhängigkeitsbeziehung zwischen dem Auftraggeber als *Principal* und dem Evaluierenden als *Agent* wird der Zustand der Unabhängigkeit automatisch verletzt, der sich dadurch charakterisiert, dass ein Akteur ohne Einwirkungen von Dritten frei handeln kann (Widmer, 2012, S. 130–131)(siehe Kapitel 2.2). Dabei können Auftraggeber als *Principals* in ihrer Interaktion mit Evaluierenden (*Agents)* auf unterschiedliche Art und Weise den Evaluationsprozess beeinflussen (siehe Kapitel 2.3) (Pleger & Hadorn, 2018; Pleger et al., 2016; Stockmann et al., 2011; Widmer, 2012, S. 136).

Interessen- und Machtasymmetrien innerhalb der Principal-Agent-Beziehung können zu Verzerrungen im Evaluationsprozess führen (Barnett & Camfield, 2016, S. 529). Um im Rahmen dieser Studie die Einflussnahme des Auftraggebers (*Principal)* auf Evaluierende

im Evaluationskontext der Vereinigten Staaten zu untersuchen, werden folglich auf die Eigenschaften des *Principals* und *Agents* und die Charakteristika dieses Beziehungsverhältnis eingegangen. Dabei sind die nachfolgend beschriebenen Annahmen der PAT zentral (Davis, Schoorman, & Donaldson, 1997; Eisenhardt, 1989; Grossman & Hart, 1983; Holmström, 1979; Jensen & Meckling, 1976; Van Slyke, 2006; Waterman & Meier, 1998; Wenger & Terberger, 1988).

Jensen und Meckling (1976, S. 308) definieren eine Agency-Beziehung als Vertrag, bei dem ein *Principal* eine andere Person (*Agent*) mit einer Aufgabe beauftragt und ihr gleichzeitig gewisse Entscheidungsbefugnisse überträgt. Diese Delegationsbeziehung, die dem *Principal* zur Realisierung seiner Interessen dient, ist geprägt durch eine asymmetrische Informationsverteilung (Ebd., 1976, S. 308). Diese Informationsasymmetrie konstatiert sich dadurch, dass der *Agent* ein Informationsvorteil hat, während dem *Principal* Informationen über den *Agent* und seine Aktivitäten fehlen (Kaluza, Dullnig, & Malle, 2003, S. 20). Die Handlungen des *Agents* beeinflussen seinen eigenen Nutzen, aber auch denjenigen des *Principals*. Durch das Vertragsverhältnis und die damit verbundene Delegation von Aufgaben und Entscheidungskompetenzen an den *Agent* profitiert der *Principal* von der spezialisierten Arbeitskraft und dem Informationsvorteil des *Agents* (Oehlrich, 2016, S. 116). Die PAT geht davon aus, dass der *Agent* mehr Informationen bezüglich der sachlichen Aufgabenbearbeitung und somit einen Informationsvorsprung gegenüber dem *Principal* hat. Dies kommt nicht ungefähr, denn gerade aufgrund der spezialisierten Fähigkeiten, Erfahrungen und Kenntnisse des *Agents* in einem Gebiet delegiert der *Principal* diesem auch die Aufgabe (Ebd., 2016, S. 121). Abhängig von der Informationsverteilung können unterschiedliche Verhaltensweisen induziert werden. Das heisst asymmetrische Informationsverteilungen können opportunistisches Verhalten begünstigen (Ebd., 2016, S. 121). Innerhalb dieser Delegationsbeziehung wird angenommen, dass beide Akteure stets im eigenen Interesse, d.h. nutzenmaximierend handeln und der *Agent* möglicherweise nicht immer im besten Interesse des *Principals* handelt (Jensen & Meckling, 1976, S. 308). Die Nutzenmaximierung der Akteure besteht aus drei Aspekten und beruht auf stabilen Präferenzen, orientiert sich an den Individualnutzen der Akteure und kann opportunistische Verhaltensweisen beinhalten (Kaluza et al., 2003, S. 20). Eigennutzenmaximierendes Verhalten tritt in unterschiedlichen Formen auf, wobei Opportunismus als stärkste Ausprägung gilt und der *Principal* grundsätzlich ein opportunistisches Verhalten seitens des *Agents* befürchten und erwarten muss. Opportunistisches Verhalten liegt bspw. vor, wenn der *Agent* vor Vertragsabschluss relevante Informationen

über seine beruflichen Qualifikationen verschweigt oder danach nicht die für den *Principal* vorteilhaftesten Aktivitäten erbringt (Oehlrich, 2016, S. 117–118). Die Akteure besitzen ein grosses Repertoire an Verhaltensweisen, wobei es zudem zur Zurückhaltung von Leistungen, eigeninteressierten Auslegung vom Vertrag oder zu Fehldarstellungen kommen kann (Kaluza et al., 2003, S. 20). Die Nutzenfunktion des *Agents* besteht aus einem grossen Repertoire an Zielen, wobei zwischen monetären Zielen wie bspw. das Honorar und nicht-monetären Zielen wie Prestige, Macht, Karriere oder Freizeit unterschieden werden kann (Ebd., 2003, S. 20). Entsprechend nimmt die PAT an, dass ein Interessenkonflikt resp. Zielkonflikt zwischen dem *Agent* und dem *Principal* existiert (Caers et al., 2006, S. 27; Waterman & Meier, 1998, S. 177). Die Aufgabendelegation an den *Agent* kann bei zweierlei opportunistischen Verhaltensweisen problematisch sein, erstens wenn wie bereits erwähnt, dem *Principal* Informationen hinsichtlich des spezifischen Leistungsverhaltens, der Motive und Handlungsmöglichkeiten des *Agents* fehlen (Oehlrich, 2016, S. 116). Diese Unsicherheit über die Eigenschaften des *Agents* nennt sich *hidden characteristics* und kann auch als Qualitätsunsicherheit bezeichnet werden, die vor Vertragsabschluss mit dem *Agent* auftritt und folglich bis zur ungeeigneten Auswahl eines Vertragspartners führen kann (Adverse selection). Als zweites Beispiel kann neben *hidden characteristics* auch *hidden intention* vorliegen, wenn der *Principal* schwierig beurteilen kann, ob der *Agent* sich an die im Vertag verankerten Inhalte während der Agency-Beziehung hält. Um dem entgegen zu wirken, kann der *Principal* versuchen die Produktivität ihrer Beziehung zu erhöhen, indem er bspw. in Fähigkeiten, Wissen und Kenntnisse des *Agents* investiert (Ebd., 2016, S. 123). Mit dem Ausmass des Informationsnachteils des *Principals* steigt somit die Gefahr, dass sich der *Agent* vordergründig an den eigenen Interessen statt der Aufgabenerfüllung im Interesse des *Principals* orientiert (Ebd., 2016, S. 116).

Die divergierenden Interessen schränkt der *Principal* einerseits durch entsprechende Anreize gegenüber dem *Agent* ein, andererseits versucht der *Principal* die Aktivitäten des *Agent* zu überwachen (Jensen & Meckling, 1976, S. 308). Das sog. Agency-Problem tritt einerseits aufgrund divergierender Ziele beider Parteien auf. Andererseits kann das Problem auftreten, wenn der *Principal* unmöglich die Arbeit vom *Agent* kontrollieren kann oder diese Kontrolle zu teuer ist (Eisenhardt, 1989, S. 58). Ohne Aufwand resp. Kosteneinsatz – sog. Agency-Kosten – ist es grundsätzlich schwierig sicherzustellen, dass der *Agent* optimale Entscheidungen im Interesse des *Principals* trifft. Die meisten Agency-Beziehungen charakterisieren sich meistens durch eine gewisse Divergenz zwischen der

Entscheidung des *Agents* und der potenziell besten Entscheidung, die den Nutzen des *Principals* maximieren würde (Jensen & Meckling, 1979, S. 308). Principal-Agent Theoretiker empfehlen verschiedene Governance-Mechanismen, um die Interessen zwischen *Principal* und *Agent* auszugleichen, wobei erstens zwischen finanziellen Anreizsystemen und zweitens kontrollierenden Governance-Strukturen unterschieden werden kann (Davis et al., 1997, S. 23). Mithilfe eines geeigneten Anreizsystems beabsichtigt der *Principal* den genannten Mechanismen entgegen zu wirken und den *Agent* zu beeinflussen. Anreizsysteme dienen dabei der Gestaltung von Principal-Agent-Beziehungen und verstehen sich gemäss Richter (1994, S. 2) als „ein auf ein bestimmtes Zielbündel abgestimmtes System von Normen" (Roiger, 2007, S. 1–2). Finanzielle Anreizsysteme mit Anreizen und Sanktionen verfolgen primär den Zweck, die unterschiedlichen Interessen zwischen *Principal* und *Agent* auszugleichen. Durch finanzielle Anreize werden *Agents* motiviert, sich gemäss den Interessen der *Principals* zu verhalten, wobei solche Anreizsysteme besonders wünschenswert sind, wenn ein signifikanter Informationsvorsprung aufseiten des *Agents* vorliegt und eine Überwachung unmöglich ist. Kontrollierende Governance-Mechanismen werden v.a. dann eingesetzt, wenn die *Principals* die Leistung der *Agents* als schlecht bewerten, was verschiedene Ursachen wie bspw. Wissensmangel, fehlende Informationen und mangelnde Fähigkeit sowie Motivation haben kann (Davis et al., 1997, S. 23–24).

Das Agency-Problem charakterisiert sich durch eine Allgemeingültigkeit, die darin besteht, ein *Agent* dazu zu bringen, sich so zu verhalten, dass er den Nutzen des Auftraggebers (*Principals*) maximiert. Diese Problemsituation und die damit verbundenen Agency-Kosten existieren in allen Organisationen und kooperativen Beziehungen (Jensen & Meckling, 1979, S. 309). So empfiehlt auch Eisenhardt (1989, S. 57) die Prinicpal-Agent-Perspektive bei Untersuchungen von Problemen mit einer kooperativen Struktur miteinzubeziehen. Aufgrund der beschriebenen Allgemeingültigkeit und der realistischen, einzigartigen und empirisch überprüfbaren Sichtweise des Agency-Modells kann das Modell für die Untersuchung von Principal-Agent-Problemen gut auf andere Kontexte adoptiert werden (Balago, 2014, S. 251). Daraus wird für vorliegende Studie abgeleitet, dass sich die PAT für die Adaption auf den Evaluationskontext eignet, wobei das Zusammenspiel zwischen Auftraggebern (*Principals*) und Evaluierenden (*Agents*) im Evaluationsprozess von einer solchen, vertraglich geregelten, kooperativen Struktur geprägt ist. Die kooperative Beziehung von Auftraggebern und Evaluierenden beginnt dabei mit der Vergabe des Evaluationsauftrags und endet bei der Fertigstellung des Evaluationsberichts. Zudem

zeigt das Ergebnis eines Reviews von unterschiedlichen empirischen Studien im Bereich des Agency-Modells, dass die PAT relevant dafür ist, dass Verhaltensproblem in Zusammenhang mit der Agency-Beziehung zu adressieren und sich das Modell zur Verbesserung von kooperativen Verhaltensweisen und somit der Performanz von *Agents* eignet. Zudem hat sich empirisch gezeigt, dass die Theorie nicht nur in profitorientierten Organisationen, sondern auch in Organisationen des NPO-Sektors angewendet werden kann (Ebd., 2014, S. 250). Auch in dieser Hinsicht beweist die Theorie eine gewisse Allgemeingültigkeit, wobei vorliegende Studie das Beziehungsverhältnis von Auftraggebern und Evaluierenden nicht vordergründig im Privatsektor, sondern v.a. auch im öffentlichen und NPO-Sektor untersucht. Weiter zeigt die Studie von Balago (2014, S. 250), dass eine Vielzahl von Agency-Modellen mit unterschiedlichen Annahmen und Limitationen bestehen. Aufgrund dieser Vielfalt werden die zentralen Annahmen für die vorliegende Studie im folgenden Abschnitt zusammenfassend dargestellt, direkt auf den Evaluationskontext übertragen und für die Untersuchung relevante Hypothesen abgeleitet.

2.5 Hypothesen und Conceptual Model

Das Ziel dieser Studie liegt in der Untersuchung wie die Unabhängigkeit von Evaluationen von Auftraggebern in den Vereinigten Staaten beurteilt wird. Dabei wird untersucht, welche Rolle die Eigenschaften von Auftraggebern und ihre Beziehung zu Evaluierenden in der Beeinflussung des Evaluationsprozesses spielen. Die zentralen Annahmen[1] der PAT werden schrittweise auf den Evaluationskontext der Vereinigten Staaten übertragen, woraus zwei Hypothesenblöcke mit insgesamt sieben Hypothesen abgeleitet werden. Die Hypothesen werden danach unterteilt, ob die unabhängige Variable das Beziehungsverhältnis zwischen den Auftraggebern und Evaluierenden oder die Eigenschaften der Auftraggeber erfasst. Die Hypothesen untersuchen dabei sowohl den Zusammenhang dieser unabhängigen Variablen zur Einflussnahme der Auftraggeber auf den Evaluationsprozess als abhängige Variablen (siehe Abbildung 3). Als Ausnahme untersucht die fünfte Hypothese wie die Vertrautheit mit Evaluationsstandards mit der Erwartung an die Unabhängigkeit von Evaluationen zusammenhängt. Die beiden Hypothesenblöcke werden in den folgenden Abschnitten nacheinander beschrieben.

[1] Die zentralen Annahmen sind nachfolgend jeweils kursiv dargestellt.

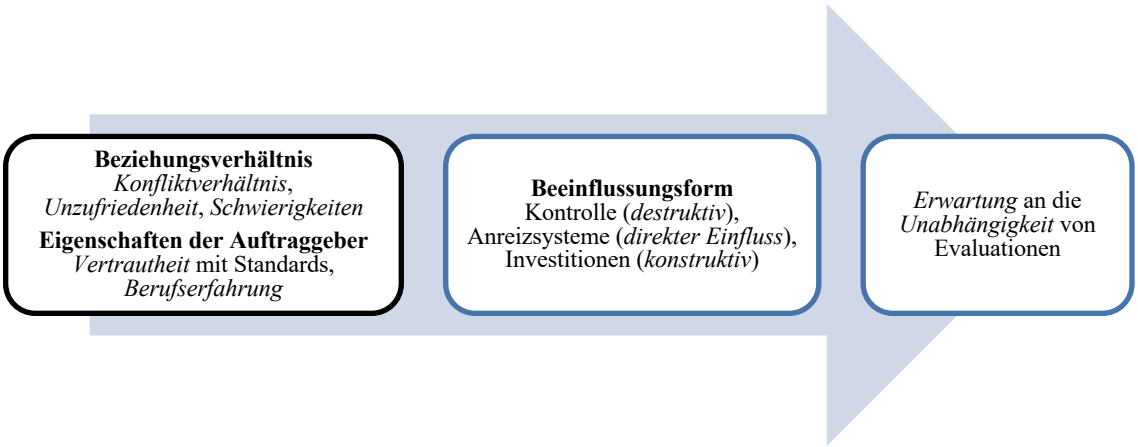

Abbildung 3: Conceptual Model (eigene Darstellung)
Anmerkung: Die unabhängigen Variablen sind im schwarzen Kasten, die abhängigen Variablen in den blauen Kästen dargestellt.

Der erste Hypothesenblock untersucht anhand der vier Hypothesen welcher Zusammenhang zwischen dem Beziehungsverhältnis der Auftraggeber und Evaluierenden und der Beeinflussungsform der Auftraggeber innerhalb des Evaluationsprozesses besteht. Dabei wird der Zusammenhang des durch den Interessenkonflikt geprägte Beziehungsverhältnisses anhand dreier Gegebenheiten – wie dem Konfliktverhältnis, der Unzufriedenheit und der Schwierigkeiten – und der Einflussnahme der Auftraggeber analysiert. Genauer wird der Zusammenhang dieser Gegebenheiten mit der destruktiven Beeinflussungsform sowie des Anreizsystems und des direkten Einflusses untersucht. Der zweite Hypothesenblock geht der Frage nach, wie die durch die Informationsasymmetrie geprägten Eigenschaften der Auftraggeber – wie die Vertrautheit mit Standards und die Berufserfahrung – einerseits mit der konstruktiven Beeinflussungsform zusammenhängen und andererseits wie die Vertrautheit mit der Erwartung an die Unabhängigkeit von Evaluationen zusammenhängt.

2.5.1 Hypothesen zum Beziehungsverhältnis zwischen Auftraggebern und Evaluierenden

Der erste Hypothesenblock untersucht den Zusammenhang des Beziehungsverhältnisses zwischen den Auftraggebern und den Evaluierenden und der Einflussnahme der Auftraggeber auf den Evaluationsprozess. Dieser Hypothesenblock besteht aus insgesamt vier Hypothesen, welche aus zwei grundsätzlichen Annahmen (siehe Abbildung 2) der PAT mit Übertragung auf den US-Evaluationskontext abgeleitet wurden. Die Annahmen fokussieren auf den Interessenkonflikt, der das Beziehungsverhältnis zwischen den Auftraggebern und Evaluierenden wesentlich prägt.

Erstens herrscht zwischen dem Principal und Agent ein Interessenkonflikt, wobei beide Akteure ihren Individualnutzen maximieren wollen (Davis et al., 1997). Für den US-Evaluationskontext kann somit abgeleitet werden, dass die Interessen von Evaluierenden grundsätzlich darin bestehen, die Evaluationsstandards und -richtlinien resp. die Guiding Principles der AEA oder die „Program Evaluation Standards" des Joint Committee on Standards zu befolgen. Entsprechend stehen ethische Verhaltensweisen, wie die Bewahrung der Unabhängigkeit von Evaluationen und die Gewährleistung der Evaluationsqualität im Vordergrund des Interesse von Evaluierenden (AEA, 2011, S. 2). Gleichzeitig verfolgen die Evaluierenden das Interesse, dass die Auftraggeber mit der Evaluation wie bspw. dem Qualitätsaspekt zufrieden sind, um nicht zuletzt an Nachfolgeaufträge zu kommen (Stockmann et al., 2011, S. 57). Bei vorhandener Zufriedenheit aufseiten der Auftraggeber sind Nachfolgeaufträge wahrscheinlicher, was wiederum dem Nutzen der Evaluierenden entspricht, den sie zu maximieren versuchen. Demgegenüber stehen die Interessen der Auftraggeber, für welche angenommen wird, dass sie ebenfalls dafür einstehen, dass die Evaluation eine hohe Qualität aufweist. Jedoch wird angenommen, dass die Auftraggeber im Vergleich zu den Evaluierenden nicht das gleiche Verständnis von Evaluationsqualität teilen. Dies wird auch daraus abgeleitet, dass bei Stiftungsratsmitgliedern als Auftraggeber von Evaluationen ein erhöhtes Potenzial identifiziert wurde, dass sie ethische Herausforderungen vergleichsweise unterschiedlich definieren als Evaluierende (Morris, 2007). Eigentlich würde es in der Natur der Evaluation liegen, dass gerade die Auftraggeber das grösste Interesse an unabhängigen Evaluationen haben, denn gerade wegen der wissenschaftlichen Integrität und Objektivität werden Evaluationen beauftragt. Die Interessen der Auftraggeber mögen aber auch darin liegen, ihre individuellen Präferenzen in den Evaluationsprozess einfliessen zu lassen, um ihren Nutzen zu maximieren (Pleger et al., 2016, S. 11). Möglichst positive Evaluationsresultate können als naheliegende Interessen der Auftraggeber angenommen werden und dienen nicht zuletzt der Legitimation ihrer Entscheide oder der Verbesserung ihres eigenen Rufs (Kaluza et al., 2003, S. 20). Dieses Interesse leitet sich bspw. aus einem bevorstehenden Finanzierungsentscheid, einer Anpassung eines politischen Programmes oder einer Politikveränderung ab. Somit wird angenommen, dass Auftraggeber entsprechend stark an den Evaluationsresultaten interessiert sind, da sich diese wiederum auf ihre Organisation, ihr Programm oder sogar die eigene persönliche Karriere – positiv oder negativ – auswirken können. Die Beziehung während dem Evaluationsprozess zwischen den Evaluierenden und den

Auftraggebern wird somit massgeblich durch dieses direkte Interesse aufseiten der Auftraggeber geprägt (Barnett & Camfield, 2016, S. 528). Aufgrund der divergierenden Interessen und der individuell angestrebten Nutzenmaximierung beider Akteure mündet das Beziehungsverhältnis unvermeidlich in einem Interessenkonflikt (Posavac, 2014, S. 121-122). Für das Beziehungsverhältnis zwischen Stiftungsratsmitgliedern und deren Stakeholdern hat Morris (2007) ebenso ein Interessenkonflikt identifiziert. Aus Auftraggebersicht bezog sich der ungelöste Interessenkonflikt mit den Evaluierenden auf das zu evaluierende Programm, wobei sich Evaluierende nicht an vereinbarte Abmachungen mit der Stiftung hielten und Besuche vor Ort nicht in Zusammenhang mit dem Evaluationsprozess standen. Weitere Interessenkonflikte standen in Zusammenhang mit der Anonymität, Vertraulichkeit und der Einverständniserklärung. Eine weitere Herausforderung bestand aus Auftraggebersicht darin, wie Ergebnisse ethisch genutzt werden, die auf eine schlechte Leistung des evaluierten Programmes hinweisen (Morris, 2007, S. 412). Daher überrascht der frühe Befund aus dem Jahr 1993 nicht, dass Konflikte bezüglich der Berichterstellung von Evaluationsresultaten als häufigstes ethisches Problem von Evaluierenden genannt wurden und die häufigsten Konflikte mit Fehldarstellung von Resultaten oder mit der Einhaltung von Offenlegungsvereinbarungen einhergehen (Morris & Cohn, 1993). Bei Uneinigkeiten dieser Art sehen sich die Auftraggeber gezwungen „[to] expend resources both in trying to instruct the agent what to do and in monitoring and policing the agent's behavior" (Mitnick, 1986, S. 4). In Zusammenhang mit solchen Uneinigkeiten, wird angenommen, dass die Auftraggeber den Evaluationsprozess und damit die Tätigkeit der Evaluierenden anhand von kontrollierenden Beeinflussungsversuchen zu steuern versuchen und entsprechende Änderungsvorschläge in Form einer destruktiven Beeinflussungsform unterbreiten. Daraus wird die erste Hypothese abgeleitet, die wie folgt lautet:

> *H1: Je stärker die Auftraggeber ein konfliktgeprägtes Verhältnis mit den Evaluierenden wahrnehmen, desto stärker ist die destruktive Beeinflussungsart der Auftraggeber ausgeprägt.*

Zudem kann angenommen werden, dass sich der Interessenkonflikt durch eine gewisse Unzufriedenheit mit Evaluationen wie der Qualität, den Resultaten, Schlussfolgerungen oder den Kompetenzen der Evaluierenden äussert. Auch Schwierigkeiten in der Zusammenarbeit mit Evaluierenden können Bestandteil des Interessenkonflikts sein, wobei diese auf unterschiedliche Ursachen zurückgeführt werden können wie bspw. mangelhafte Qualität, Motivation, Kompetenzen, Ressourcen oder fehlendes Verständnis für die

zu evaluierende Organisation (Pleger & Hadorn, 2018). Sowohl die Unzufriedenheit der Auftraggeber als auch die Schwierigkeiten in der Zusammenarbeit mögen, wie bei der ersten Hypothese auch, zu kontrollierenden Beeinflussungsversuchen aufseiten der Auftraggeber führen, die sich in einer destruktiven Beeinflussungsform konstatieren. Die zweite und dritte Hypothese untersuchen daher folgende positive Zusammenhänge:

> *H2: Je stärker die Auftraggeber mit einer in Auftrag gegebenen Evaluation unzufrieden sind, desto stärker ist die destruktive Beeinflussungsart der Auftraggeber ausgeprägt.*

> *H3: Je mehr Schwierigkeiten die Auftraggeber in der Zusammenarbeit mit den Evaluierenden wahrnehmen, desto stärker ist die destruktive Beeinflussungsart der Auftraggeber ausgeprägt.*

Zweitens versucht der Principal anhand von Kontrollmechanismen wie Sanktionen oder Anreizen das Verhalten des Agents an die Interessen des Principals anzugleichen und somit das Risiko des eigennützigen Verhaltens des Agents zu minimieren (Van Slyke, 2006). Gemäss der PAT versuchen Auftraggeber einerseits die Aktivitäten der *Agents* zu überwachen (siehe *H1-H3)*, andererseits schränken Auftraggeber divergierende Interessen durch entsprechende Anreize gegenüber Evaluierenden ein (Jensen & Meckling, 1976, S. 308). Mithilfe eines geeigneten Anreizsystems versuchen die Auftraggeber dem Interessenkonflikt entgegenzuwirken und die Evaluierenden zu beeinflussen (Roiger, 2007, S. 1). In vorliegender Studie wird unter dem Anreizsystem, die im Rahmen des BUSD-Modells beschriebene Druckausübung aufseiten der Auftraggeber verstanden, die sich auf der Dimension des Entfaltungsgrads der Beeinflussung (*Explicitness of influence*) einordnen lässt und in Anreize und Sanktionen unterteilt werden kann. Die erste Dimension der Beeinflussungsintention (*Direction of influence*) erhält in Zusammenhang des Anreizsystems eine zweitrangige Bedeutung, da vorliegend angenommen wird, dass Anreizsysteme per se destruktiv sind. Dies resultiert einerseits aus einer dem BUSD-Modell zugrundeliegender impliziten Annahme und gleichzeitigen Limitation des Modells, dass die Arbeit der Evaluierenden fundamentalen Evaluationsstandards und der wissenschaftlichen Integrität folgt. Somit stellen Anreizsysteme Beeinflussungsversuche seitens der Auftraggeber dar, bspw. die Evaluationsresultate entgegen der Interessen der Evaluierenden, die in der Aufrechterhaltung von Evaluationsstandards liegen, zu beeinflussen. Andererseits können Verzerrungen von Evaluationsresultaten besonders dann auftreten,

wenn diese an Zahlungen gekoppelt werden und bei divergierenden Interessen gar finanzielle Konsequenzen zur Folge haben können (Barnett & Camfield, 2016, S. 528-529). Dieser finanzielle Anreiz verdeutlicht die destruktive Art der Einflussnahme eines Anreizsystems. Folglich fällt eine Unterscheidung zwischen destruktiver und konstruktiver Einflussnahme weg, wobei eine graduelle Abstufung innerhalb der destruktiven Einflussnahme durchaus sinnvoll wäre, jedoch nicht Bestandteil dieser Forschung ist. Demgegenüber kann in Zusammenhang mit Anreizsystemen anhand des Modells zwischen der direkten, offensichtlichen (expliziten) und der indirekten, subtilen (impliziten) Druckausübung auf den Evaluierenden unterschieden werden, die nicht nur den Evaluierenden, sondern mitunter auch die Unabhängigkeit von Evaluationen beeinflusst (Pleger & Sager, 2018, S. 169). Diese Unterscheidung kann direkt aus der Art des Anreizsystems abgeleitet werden. Eine explizite Druckausübung in Form eines Anreizes oder einer Sanktion liegt vor, wenn direkte, konkrete Änderungsaufforderungen hinsichtlich der Evaluation oder den Evaluationsbericht geäussert werden. Bei der impliziten Druckausübung werden keine konkreten Änderungsaufforderungen geäussert, sondern nur die Konsequenzen eines Nichtbefolgens impliziter Forderungen betont. Eine explizite, direkte Druckausübung in Form eines Anreizes liegt bspw. vor, wenn die Auftraggeber den Evaluierenden klare Anreize für Ergebnissänderungen nahelegen, indem auf Nachfolgeaufträge hingewiesen wird. Stellen Auftraggeber den Evaluierenden bspw. in Aussicht, dass der Evaluationsbericht nicht veröffentlicht oder das Honorar nicht bezahlt wird, handelt es sich um eine implizite, indirekte Durckausübung in Form einer Sanktion (Pleger & Hadorn, 2018). Für einen Interessenkonflikt gibt es wie bereits erwähnt viele mögliche Gründe (Morris, 2007; Morris & Cohn, 1993). Unabhängig der spezifischen Konfliktgründe steht im Forschungsinteresse, wie das Konfliktverhältnis zwischen Auftraggebern und Evaluierenden mit dem Anreizsystem und dem direkten Einfluss aufseiten der Auftraggeber zusammenhängt. Daraus abgeleitet, lautet die vierte zweigeteilte Hypothese wie folgt:

> *H4: Je stärker die Auftraggeber ein konfliktgeprägtes Verhältnis mit den Evaluierenden wahrnehmen, a) desto häufiger setzen die Auftraggeber Anreizsysteme ein und b) desto häufiger üben die Auftraggeber direkten, expliziten Einfluss auf den Evaluationsprozess aus.*

2.5.2 Hypothesen zu den Eigenschaften der Auftraggeber

Der zweite Hypothesenblock untersucht den Zusammenhang der Eigenschaften der Auftraggeber mit einerseits der konstruktiven Einflussnahme der Auftraggeber auf den Evaluationsprozess und andererseits der Erwartung an die Unabhängigkeit von Evaluationen.

Dieser Hypothesenblock besteht aus drei Hypothesen, welche analog zum ersten Abschnitt aus zwei grundsätzlichen Annahmen der PAT mit Übertragung auf den US-Evaluationskontext abgeleitet wurden. Die Annahmen fokussieren auf der Informationsasymmetrie, welche sich sowohl in den Eigenschaften der Auftraggeber als auch der Evaluierenden niederschlägt.

Drittens existiert eine Informationsasymmetrie zwischen beiden Akteuren, wobei der Principal über weniger Informationen als der Agent verfügt (Kaluza et al., 2003, S. 20). Es kann davon ausgegangen werden, dass die Auftraggeber weniger Informationen in Bezug auf den Evaluationsprozess und die damit verbundene Evaluationstätigkeit der Evaluierenden verfügen. Dabei existiert eine Informationsasymmetrie zulasten der Auftraggeber, weil die Evaluierenden mehr Knowhow, Fähigkeiten und Erfahrungen im Bereich der sachlichen Evaluationsdurchführung haben (Oehlrich, 2016, S. 121). Im US-Evaluationskontext kann davon ausgegangen werden, dass die Kenntnis und Vertrautheit der „Program Evaluation Standards" bei den Evaluierenden tendenziell stärker ausgeprägt sind als bei Auftraggebern. Diese Annahme rührt daher, dass die Auftraggeber möglicherweise nicht die gleiche professionelle Sozialisation bei der Durchführung von Evaluationen durchlaufen haben wie Evaluierende und entsprechend ein anderes Verständnis von ethischen Evaluationen haben (Morris, 2007, S. 413). Daraus wird vorliegend abgeleitet, dass die Divergenz der Kenntnis und Vertrautheit der Evaluationsstandards zwischen den Auftraggebern und Evaluierenden Aufschluss darüber gibt, wie unterschiedlich sie ethische Standards wahrnehmen und welches Qualitätsverständnis bezüglich der Evaluation vorherrscht. Brown und Newman (1992, S. 665–661) zeigen, dass die Berufserfahrung und Bildung zu einem gemeinsamen Evaluationsverständnis zwischen Evaluierenden und Auftraggebern führen und dieses gemeinsame Verständnis von entsprechender Einigkeit bezüglich erwarteter Verhaltensweisen zeugt. V.a. bei Auftraggebern von Evaluationen und den Personen mit keiner oder mässiger Evaluationskenntnis wurde wenig Übereinstimmung hinsichtlich des Evaluationsverständnisses gefunden. Daraus wurde abgeleitet, dass diese Anspruchsgruppen im Evaluationsbereich ausgebildet werden sollen, um mitunter zu erfahren was sie von Evaluierenden erwarten sollen. Folglich wird vorliegend angenommen, dass die Vertrautheit mit Evaluationsstandards von einem besseren Evaluationsverständnis zeugt und die Erwartungen hinsichtlich der ethischen Verhaltensweisen resp. der Unabhängigkeit von Evaluationen entsprechend höher ausfallen, als bei Personen ohne Vertrautheit mit Evaluationsstandards.

Die Informationsasymmetrie zwischen dem *Principal* und *Agent* wird durch die Vertrautheit mit Evaluationsstandards repräsentiert, wobei für die vorliegende Studie der positive Zusammenhang einer kleineren Informationsasymmetrie aufgrund einer höheren Vertrautheit von Evaluationsstandards und der Erwartungen an die Unabhängigkeit von Evaluationen untersucht wird. Die fünfte Hypothese lautet wie folgt:

> *H5: Je mehr Auftraggeber mit den Evaluationsstandards vertraut sind, desto höher sind die Erwartungen an die Unabhängigkeit von Evaluationen der Auftraggeber ausgeprägt.*

Für die Informationsasymmetrie zwischen den Auftraggebern und den Evaluierenden, die sich in der ungleichen Vertrautheit mit Evaluationsstandards konstituieren kann, wird angenommen, dass sich diese nicht nur in divergierenden Erwartungen an Evaluationen niederschlägt, sondern auch mit der Art wie die Auftraggeber den Evaluationsprozess beeinflussen, zusammenhängt. Sind Auftraggeber mit den Evaluationsstandards vertraut – d.h. es herrscht eine relativ kleinere Informationsasymmetrie – wird vorliegend angenommen, dass dies in positivem Zusammenhang mit einer konstruktiven Beeinflussungsform steht, welche den Evaluationsprozess bereichert. Der Befund, dass die ethische Sensitivität von Evaluierenden u.a. mit der Kenntnis der vorgeschriebenen Normen bezüglich ethischer Fragen zusammenhängt, zeugt von einem ähnlichen Zusammenhang (Desautels & Jacob, 2012). Übertragen auf die Auftraggeberperspektive kann davon abgeleitet angenommen werden, dass die Vertrautheit mit Evaluationsstandards auch seitens der Auftraggeber eine gewisse Sensitivität gegenüber ethischen Dilemmata impliziert. Brown und Newman (1992, S. 661) nehmen an, dass gebildete Auftraggeber gegenüber unethischen Verhaltensweisen sensibilisiert sind und dabei eine gesunde Austauschbeziehung während dem Evaluationsprozess pflegen. Die Bildung bezüglich ethischer Standards schlägt sich gewissermassen in der Vertrautheit mit den Evaluationsstandards nieder. Diese Vertrautheit kann sich folglich in einer gesunden Austauschbeziehung in Form einer konstruktiven Einflussnahme auf den Evaluationsprozess äussern. Die Beeinflussungsform kann als konstruktiv eingeordnet werden, wenn die Einflussabsicht der Evaluationsqualität zugutekommt und der evidenzbasierten Forschung nützlich ist (siehe Kapitel 2.3) (Pleger & Sager, 2018, S. 169). Zur Untersuchung dieses positiven Zusammenhangs lautet die sechste Hypothese wie folgt:

> *H6: Je mehr Auftraggeber mit den Evaluationsstandards vertraut sind, desto stärker ist die konstruktive Beeinflussungsart der Auftraggeber ausgeprägt.*

Viertens wird der Informationsvorteil vom Agent zu seinen Gunsten ausgenützt, wobei er eigennützige Ziele verfolgt, ohne im Interesse des Principals zu handeln (Eisenhardt, 1989). Die Informationsasymmetrie zwischen den Auftraggebern und Evaluierenden kann sich darin konstatieren, dass den Auftraggebern jegliche Informationen bezüglich der spezifischen Qualifikation, des Wissens, der Motive und Handlungsmöglichkeiten sowie des konkreten Leistungsverhaltens der Evaluierenden auf dem relevanten Gebiet fehlen (*hidden characteristics*). Obwohl diese Situation den Zeitpunkt vor Vertragsabschluss betrifft, ist sie genauso relevant für den vollständigen Evaluationsprozess, der in dieser Studie untersucht wird (Oehlrich, 2016, S. 123). Während Evaluierende die Rolle der „fachkundigen Gutachter" einnehmen, ist es für die Auftraggeber riskant sich auf die Professionalität eines einzelnen Evaluierenden zu verlassen. Dieses Risiko erwächst u. a. aus dem Evaluationskontext heraus, wo viele Belastungen wie bspw. mangelnde zeitliche Ressourcen auftreten und Evaluierende in ihrer Evaluationstätigkeit behindern können (Barnett & Camfield, 2016, S. 531). Für diese Informationsasymmetrie zulasten der Auftraggeber wird angenommen, dass diese noch stärker ausgeprägt ist, wenn die Auftraggeber relativ weniger Evaluationserfahrung aufweisen. Vergeben Auftraggeber schon seit vielen Jahren Evaluationsaufträge und haben in ihrer Karriere über ein Duzend Evaluationen durchgeführt, sind den Auftraggebern Eigenschaften der Evaluierenden durch vergangene Auftragsverhältnisse womöglich eher bekannt, oder können durch ihre Erfahrung die fehlenden Informationen eher abschätzen als dies bei „Auftraggeber-Anfängern" der Fall wäre. Fehlen den Auftraggebern die Berufserfahrung, können diese vor Vertragsabschluss nur schwierig beurteilen, ob sich die Evaluierenden an die Vertragsinhalte halten (*hidden intention*)(Oehlrich, 2016, S. 123). Je grösser das Informationsdefizit der Auftraggeber ist, desto grösser ist auch die Gefahr, dass sich die Evaluierenden primär an ihren Interessen statt der Aufgabenerfüllung im Interesse der Auftraggeber orientieren (Ebd., 2016, S. 116). Durch Investitionen in Fähigkeiten, Wissen und Kenntnisse der Evaluierenden während des Evaluationprozesses können Auftraggeber jedoch diesem Informationsnachteil entgegenwirken und damit die Qualität der Evaluation verbessern (Ebd., 2016, S. 123). Diese Investitionen finden in Form einer positiven, konstruktiven Einflussnahme der Auftraggeber auf den Evaluationsprozess statt, welche die Evaluationsqualität erhöht (Pleger & Sager, 2018). Diese Annahme impliziert, dass die Situationen der *hidden characteristics* und *hidden intention* insofern erweitert werden, dass sie sich nicht „ex ante" (Zaggl, 2012), sondern „ex post" auf den Zeitpunkt nach Vertragsabschluss resp. auf den Evaluationsprozess beziehen und die Auftraggeber mittels aktiver

Investitionen die daraus resultierenden negativen Konsequenzen zu minimieren versuchen (Oehlrich, 2016). Die siebte Hypothese lautet daher wie folgt:

> *H7: Je weniger Berufserfahrung die Auftraggeber in ihrer Tätigkeit aufweisen, desto stärker ist die konstruktive Beeinflussungsart der Auftraggeber ausgeprägt.*

2.5.3 Forschungsfragen

Die Studie untersucht *wie die Unabhängigkeit von Evaluationen in den USA beurteilt wird* und *welche Rolle die Eigenschaften der Auftraggeber und ihre Beziehung zu Evaluierenden in der Beeinflussung des Evaluationsprozesses spielen.* Dazu lassen sich nachfolgende Forschungsfragen ableiten, die zur Beantwortung dieser übergeordneten Forschungsfrage dienen.

- Inwiefern hängt ein von den Auftraggebern als konfliktgeprägt wahrgenommenes Evaluationsverhältnis mit ihrer Einflussnahme auf den Evaluationsprozess zusammen? Wie werden divergierende Interessen seitens der Auftraggeber ausgeglichen?
- Wie hängt die negative Wahrnehmung der Auftraggeber gegenüber Evaluierenden mit ihrer Einflussnahme auf den Evaluationsprozess zusammen?
- Gibt es einen Zusammenhang zwischen der Vertrautheit von Auftraggebern mit Evaluationsstandards und deren Erwartung an die Unabhängigkeit von Evaluationen?
- Welche Rolle spielt die Berufserfahrung und die Vertrautheit mit Evaluationsstandards in der Einflussnahme der Auftraggeber auf den Evaluationsprozess?
- Wie beurteilen Auftraggeber die Wichtigkeit von unabhängigen Evaluationen und wie nehmen Auftraggeber die Unabhängigkeit von Evaluationen sowie ihre eigene Einflussstärke wahr?
- Was sind von Auftraggebern wahrgenommene Gründe für das konfliktgeprägte Verhältnis und welche präventiven Massnahmen werden vorgeschlagen, um ein fruchtbares Umfeld für aussagekräftige Evaluationen zu schaffen?

Im Ländervergleich wird übergeordnet die Frage nach den Unterschieden und Gemeinsamkeiten der Auftraggeber der USA und der Schweiz in Bezug auf individuelle Eigenschaften sowie ihrem Beziehungsverhältnis zu Evaluierenden untersucht. Dabei werden die Erwartungen der Auftraggeber beider Länder gegenüber der Unabhängigkeit von Evaluationen sowie die Kenntnis, Vertrautheit und Wichtigkeit von nationalen Evaluationsstandards verglichen. Zudem wird untersucht, ob Unterschiede in der Reaktion oder Unterstellung von Änderungsvorschlägen in beiden Ländern existieren. Abschliessend

werden wahrgenommene Schwierigkeiten in der Zusammenarbeit mit Evaluierenden sowie vorgeschlagene präventive Massnahmen einander gegenübergestellt. Das der Untersuchung zugrundeliegende Forschungsdesign wird im nächsten Kapitel beschrieben.

3 Forschungsdesign

In folgendem Kapitel wird das Forschungsdesign der Studie beschrieben und genauer auf das Datenerhebungsinstrument, das Material und Sample sowie die Untersuchungseinheiten eingegangen. In einem nächsten Abschnitt wird die Fragebogenentwicklung mitsamt Fragebogenaufbau und Pretest beschrieben. Weiter wird auf die Operationalisierung der Konstrukte eingegangen, welche zur Messung der Hypothesen gebildet wurden. In einem letzten Schritt werden die Datenauswertungsmethoden vorgestellt.

Für die Überprüfung der angenommenen Effekte werden die theoretisch hergeleiteten Hypothesen mittels einer Online-Befragung der Auftraggeber von Evaluationen in den USA getestet. Die Erhebung der wahrgenommenen Beeinflussungsform von Auftraggebern – ob konstruktiv oder destruktiv – stellt das Zentrum der Untersuchung dar. Neben der Ermittlung von Fakten, Wissen und Meinungen werden somit vordergründig Bewertungen erhoben, wofür die Befragung ein unverzichtbares Erhebungsinstrument darstellt (Schnell, Hill, & Esser, 2018). Die Befragung birgt gegenüber Face-to-Face-Interviews eine Reihe von Vorteilen wie bspw. die Möglichkeit aufseiten der Befragten die Fragen besser durchzudenken, die mögliche Beeinflussung von Interviewern fällt weg und sie sind kostengünstiger. Wichtig zu beachten ist, dass bei Verständnisproblemen keine Hilfestellung erfolgen kann und hohe Anforderungen an die selbsterklärende und einfache Fragebogengestaltung gestellt werden (Diekmann, 2013, S. 514). In vorliegender Studie wurde eine internetgestützte Befragung in Form eines „Web-Surveys“ durchgeführt, wobei der Fragebogen anhand der Umfragesoftware von Qualtrics ausgeführt und über E-Mailverteilungen verbreitet wurde (Schnell et al., 2018, S. 343). Mit der cloudbasierten Umfragetechnologie von Qualtrics wurde die Online-Befragungen gestaltet, gesendet und fortlaufend analysiert (Qualtrics, 2019b). Als Spezialfall der schriftlichen Befragung biete die Online-Befragung vielfältige Vorteile. Internetgestützte Befragungen können schnell durchgeführt werden, wobei die Daten unmittelbar abgespeichert werden. Während der Befragungsdauer werden dadurch zeitnahe ermöglicht. Neben der genauen Bestimmung der Fragenreihenfolge können in Abhängigkeit der Antworten Sprünge anhand von Filterfragen programmiert werden, sodass nur für die Respondenten relevante Fragen

angezeigt werden. Das Befragtenverhalten kann zudem aufgezeichnet werden, wobei unterschiedliche Reaktionen der Respondenten beim Ausfüllen des Online-Fragebogens registriert werden können, was wiederum Rückschlüsse auf Schwierigkeiten innerhalb des Online-Fragebogens erlaubt (Diekmann, 2013, S. 522–523). Demgegenüber birgt die Befragung gewisse Nachteile, die sich durch unterschiedliche Fehlerquellen wie Befragten- oder Fragemerkmale charakterisieren. Befragtenmerkmale können Probleme wie die soziale Erwünschtheit, das Response-Set sowie das Problem der Non-Attitude hervorrufen (Ebd., 2013, S. 447). Diese möglichen Fehlerquellen wurden sowohl bei der Fragebogen- und Skalenentwicklung und der Interpretation der Resultate berücksichtigt. Im nächsten Abschnitt wird der Untersuchung zugrundeliegende Material und Sample vorgestellt.

3.1 Material und Sample

Die Studie beschränkt sich auf die Untersuchung der Auftraggeberperspektive im Evaluationskontext der USA. Der US-Evaluationskontext eignet sich als Untersuchungsgegenstand, da auch die USA neben der Schweiz eine fortschrittliche Evaluationskultur aufweisen (Pleger & Hadorn, 2018). Das Zielsample setzt sich aus allen US-Auftraggebern von Evaluationen zusammen, die momentan oder in der Vergangenheit Evaluationen durchgeführt haben. Genauer fungieren diejenigen Auftraggeber als Untersuchungseinheiten, die in verschiedenen Sektoren wie dem privaten, öffentlichen oder NPO-Sektor sowie an Universitäten oder Hochschulen tätig sind.

Wie Pleger und Hadorn (2018, S. 6) zeigen, arbeiten fast alle befragten Auftraggeber im öffentlichen Sektor (92%), wobei lediglich 6% in NPO oder selbstständig (2%) tätig sind. Diese Verteilungsangaben fungieren als Orientierungsrahmen wie die US-Auftraggeber nach Sektoren verteilt sein können, wobei die Bedeutung vom öffentlichen Sektor, aber auch von Universitäten und Hochschulen für die vorliegende Studie abgeleitet angenommen und berücksichtigt wird. Für die Rekrutierung von Respondenten wurden im Vorfeld E-Mailadressen von Personen gesammelt, die im öffentlichen, privaten und NPO-Sektor sowie an Universitäten oder Hochschulen tätig sind. Relevante E-Mailadressen von potentiellen Auftraggebern aus den unterschiedlichen Sektoren wurden systematisch online identifiziert. Um auf der Ebene der Institutionen die für die Untersuchung relevanten Einheiten auszuwählen, wurde nach festen Regeln vorgegangen, was einer bewussten Auswahl entspricht (Schnell et al., 2018, S. 244). Innerhalb der durch die bewusste Auswahl selektierten Institutionen eines jeweiligen Sektors wurde darauf geachtet, dass möglichst offizielle E-Mailadressen gesammelt werden, die nicht von einer privaten E-Maildomain

wie bspw. Gmail stammen (TheWindowsClub, 2015). Weiter wurde keine zu enge Auswahl an E-Mailadressen von möglichen Respondenten getroffen, um eine mögliche Verzerrung durch diese Selektion zu vermeiden. Zur systematischen Suche von E-Mailadressen im öffentlichen Sektor wurde die offizielle Webseite der amerikanischen Regierung verwendet, welche für jeden Staat und dessen Behörden Websitelinks und Kontaktinformationen bereitstellt (usa.gov, 2019). Die Angaben zu den jeweiligen Staaten wurden in alphabetischer Reihenfolge durchgegangen, wobei zunächst auf der generellen Webseite des *State Governments* und danach die untergeordneten Webseiten der *State Agencies* abgesucht wurden. Sowohl für die Universitäten und Hochschulen als auch den Privat- und NPO-Sektor kann die bewusste Auswahl von relevanten Webseiten weiter in eine Auswahl nach dem Konzentrationsprinzip differenziert werden. Dieses Auswahlverfahren zeichnet sich dadurch aus, dass diejenigen Fälle selektiert werden, bei denen ein interessierendes Merkmal stark ausgeprägt ist, sodass beinahe die ganze Verteilung in der Grundgesamtheit durch diese Fälle bestimmt ist (Schnell et al., 2018, S. 273). In vorliegender Arbeit fungiert das Auswahlverfahren jedoch nicht zur Selektion der Fälle, die durch die individuellen E-Mailadressen erreicht werden, sondern auf einer übergeordneten Ebene zur Auswahl der Webseiten auf dessen Basis wiederum die Fälle selektiert werden. Das interessierende Merkmal für die Auswahl nach dem Konzentrationsprinzip konstituiert sich jeweils aus einem sektorspezifischen Ranking, wobei innerhalb eines gewissen Ranges die Webseiten der *besten* Universitäten und Hochschulen, der *grössten* gemeinnützigen Organisationen und der *umsatzstärksten* Firmen selektiert wurden. Auf dieses Auswahlverfahren wurde aus Gründen knapper zeitlicher und finanzieller Ressourcen zurückgegriffen, um eine Einschränkung der Webseiten zu erfahren, die zur Fallselektion dient. Anhand des *QS World University Rankings* fürs 2019 wurden E-Mailadressen von Universitäten und Hochschulen zwischen dem ersten und 90. Rang zusammengesucht (Top Universities, 2018). Dabei wurden jeweils die allgemeinen Info-E-Mailadressen und diejenigen von Mitarbeitenden aus Fakultäten und Departementen diverser Bereiche wie der Politik und Policy, Gesundheit, Entwicklung, Bildung, dem Verwaltungs- und NPO-Management sowie von Forschungsabteilungen gesammelt. Für den Privatsektor diente das *Fortune 500* Ranking der systematischen Auswahl von Firmenwebsites, aus welchen in einem nächsten Schritt E-Mailadressen selektiert wurden. Das Ranking umfasst die 500 Firmen, die zwei Drittel des amerikanischen BIP repräsentieren (Fortune, 2019). Insgesamt wurden E-Mailadressen der Firmen mit Sitz in den USA herausgesucht, welche dem ersten bis 100. Rang zugeteilt sind. Das Forbes Ranking *The 100*

Largest U.S. Charities diente der Suche von E-Mailadressen im NPO-Sektor, wobei alle 100 gemeinnützigen Organisationen durchsucht wurden (Forbes, 2019). Insgesamt wurden daraus 55‘683 E-Mailadressen gesammelt, wovon 54 Prozent dem öffentlichen Sektor, 35 Prozent den Universitäten und Hochschulen, 8 Prozent dem NPO-Sektor und 4 Prozent dem Privatsektor zugeordnet werden konnten (siehe Tabelle 1).

Tabelle 1: E-Mailverteilungen in Qualtrics

	Total E-Mails	04.04.19 8:03 AM*	11.04.19 0:54 PM	24.04.19 8:15 AM	25.04.19 7:20 AM	29.04.19 9:05 AM	02.05.19 8:37 AM
		1. Teil	2. Teil	3. Teil Reminder	4. Teil	5. Teil	6. Teil Reminder
Universität/ Hochschulen	19‘291 (34.6%)	8’740	10’051	-	500	-	19’163
Nonprofit-Sektor	4’179 (7.5%)	4’179	-	-	-	-	-
Privatsektor	2’326 (4.2%)	2’326	-	-	-	-	-
Öffentlicher Sektor	29’887 (53.7%)	8’755	14’947	-	1’000	5’185	-
Total	55’683 (100%)						
AEA	1’000	1’000	-	999	-	-	-
fehlgeschlagen		*11*	*9*	*-*	*-*	*2*	*-*
nicht zustellbar		*4’633*	*1’059*	*15*	*55*	*180*	*385*
Total	76’845	25’000	24’998	999	1’500	5’185	19’163

*Anmerkung: Die Zeitangaben sind auf die amerikanische Zeit (GMT-4) in Washington, D.C., District of Columbia, USA umgerechnet.

Im Rahmen des genehmigten Forschungsantrags bei der AEA konnte der Link zur Online-Befragung einem Sample von 1‘000 E-Mailadressen der AEA-Mitglieder gesendet werden (AEA, 2019a). Um die höchstmögliche Anonymität zu gewährleisten, wurde der Umfragelink anonymisiert und in der E-Maileinladung integriert. Im Gegensatz zu einem individualisierten Link, lässt der anonyme Link weder Rückschlüsse über die Grundgesamtheit noch über die Rücklaufquote zu (Qualtrics, 2019b). Online-Befragungen sind besonders für spezielle Populationen mit Internetzugang geeignet, für welche eine Liste mit E-Mailadressen existiert. Dieser Typ der listenbasierten Online-Befragung basiert normalerweise zwar auf einer Zufallsauswahl oder Vollerhebung, beschreibt die Vorgehensweise für die Studie mit dem beschriebenen Zielsample als spezielle Population jedoch am besten. Die E-Mailverteilung an die AEA kommt diesem Typ am nächsten, da die Sampleziehung durch die AEA möglicherweise zufällig erfolgte. Bei den restlichen E-Mailverteilungen liegt keine Zufallsauswahl vor, jedoch eine Liste von möglichst vielen Personen der jeweiligen Institutionen aus dem öffentlichen, privaten und NPO-Sektor (Diekmann, 2013, S. 528).

Die Datenerhebung wurde am 4. April 2019 mit der ersten E-Mailversteilung über Qualtrics gestartet und am 7. Mai 2019 beendet. Der Untersuchungszeitraum betrug 34 Tage und setzte sich aus einer sechsteiligen E-Mailverteilung über Qualtrics zusammen (siehe Tabelle 1). Aufgrund der Begrenzung von Qualtrics, dass wöchentlich jeweils maximal 25‘000 E-Mailadressen angeschrieben werden können, setzte sich der 1. Teil der E-Mailverteilung aus insgesamt 25‘000 E-Mailadressen aus allen Sektoren zusammen, wovon sich 1‘000 E-Mails an das Sample der AEA-Mitglieder richteten. Am 11. April folgte der 2. Teil der E-Mailverteilung (24‘998 E-Mails), mittels welcher Personen aus Universitäten und Hochschulen sowie dem öffentlichen Sektor angeschrieben wurden. Trotz der hohen Anzahl an versendeten E-Mails war der Rücklauf innerhalb der ersten drei Wochen gering. Als Ursachen wird vermutet, dass viele der versendeten E-Mails im Spamordner der jeweiligen Organisation landeten oder blockiert wurden, was sogar bei einer Test-Email an die ZHAW-E-Mailadresse der Fall war. Gemäss Qualtrics (2019a) kann es sein, dass die Qualtrics-Domain, von welcher die E-Mails verteilt wurden, nicht auf der Whitelist der Organisationen stehen. Eine weitere Ursache könnte in der fehlenden Motivation, Zeit oder dem Interesse zur Teilnahme der angeschriebenen Personen liegen. Einzelner Rückfragen bezüglich der Umfrageeinladung kann entnommen werden, dass der Begriff „evaluation client“ nicht verstanden wurde, was wiederum als mögliche Ursache für die tiefe Teilnahme gedeutet wurde. Daraus abgeleitet wurden Massnahmen zur Verbesserung der Teilnehmerrekrutierung eingeleitet. Erstens wurden nicht nur E-Mails über die Verteilungsmöglichkeit von Qualtrics versendet, sondern ebenso von privaten E-Mail-Domains wie *Me* und *Gmail*. Weitere Recherchen zur Ursache der Spammarkierung von E-Mails haben u.a. ergeben, dass der E-Mail-Betreff keine speziellen Zeichen[2] enthalten sollte (Webengage, 2019). Gemäss einem Test zeigten diese Massnahmen Wirkung und es wurden weniger E-Mails blockiert. Um die Aufmerksamkeit, Dringlichkeit sowie das Interesse und die Motivation zur Teilnahme zu erhöhen, wurden die Personen mit persönlicher Ansprache angeschrieben. Obwohl das Angebot einer Verlosung am Fragebogenende die Rücklaufquote positiv beeinflussen kann, wurde absichtlich darauf verzichtet, um keine falschen Anreize zu setzen oder ein Selection Bias hervorzurufen (Diekmann, 2013, S. 528). Konnte eine Institution nur via Kontaktformular erreicht werden, wurde diese Kontaktmöglichkeit genutzt. Drittens wurde in Anbetracht

[2] Folglich wurde der vorher mit einem Ausrufezeichen enthaltene Betreff, der v.a. die Aufmerksamkeit erwecken sollte *(„Your Participation Counts! Survey on the Independence of Evaluations“),* abgeändert zu *„Inquiry of support for survey of commissioners of evaluation“*.

der Rückmeldungen die E-Maileinladung textlich angepasst, sodass der scheinbar verständlichere Begriff „evaluation commissioner“ ebenfalls im Text integriert wurde. Weiter wurde im Text stärker hervorgehoben, wer genau an der Umfrage teilnehmen kann und dass das Weiterleiten des Umfragelinks erwünscht wird. Für mehr Hintergrundinformationen zur Forschung wurde auf die Studie von Pleger und Hadorn (2018) verwiesen. Für diesen ergänzenden Versand, der vom 11. April bis zum 7. Mai dauerte, wurden zusätzliche 2‘853 E-Mailadressen gesammelt. Bei der persönlichen Anschrift wurde generell darauf geachtet, dass Personen mit einem Titel wie *Director, Executive Director, President, CEO, Programm Director, Project Director, Grants Officer* oder ähnliches angeschrieben wurden (Morris, 2007, S. 410), der Titel auf eine Evaluationstätigkeit hinwies oder darin die Begriffe *Evaluation* oder *Research* enthalten waren. Einerseits wurden die Leiter von Evaluationseinheiten verschiedener Programme, Funds und Organisationseinheiten sowie Fachorganisationen der United Nations (UN) angeschrieben, die auf der Website der UN gelistet sind (United Nations, 2019). Andererseits wurde die Webseite des *Council on Foundations* (2019) dazu genutzt, weitere Stiftungen systematisch anzuschreiben. Die für Nicht-Mitglieder zugängliche alphabetische Auflistung von US-Stiftungen diente dieser Vorgehensweise (siehe auch Morris, 2007), wobei alle Stiftungen bis zur *The San Fransisco Foundation* kontaktiert wurden. Dem Ratschlag von Thomas Schwandt folgend (via E-Mailkorrespondenz) hat die XCeval Gruppe des Evaluation Portals (Balzer, 2019) die Weiterleitung der Umfrageanfrage an die Mitglieder unterstützt. Gewisse Universitäten sind zudem aufgrund ihres universitären Programmangebots im Bereich Evaluation auf der AEA-Website gelistet. Als zusätzliche Massnahme wurden von diesen Universitäten weitere Personen mit einer Evaluationstätigkeit angeschrieben (AEA, 2019b). Weiter wurde der Umfragelink über die Social Media Verteilungsfunktion von Qualtrics auf Linkedin und Facebook über das persönliche Profil der Autorin im eigenen Netzwerk verbreitet (Qualtrics, 2019d).

Neben diesen zusätzlichen Rekrutierungsmassnahmen wurde die geplante sechsteilige E-Mailverteilung via Qualtrics fortgesetzt. Am 24. April wurde ein Friendly-Reminder an die AEA-Mitglieder versendet (3. Teil der E-Mailverteilung). Im 4. Teil der E-Mailverteilung wurden am 25. April insgesamt 1'500 E-Mails an Universitäten, Hochschulen und den öffentlichen Sektor versendet. Im Rahmen des 5. und 6. Teils der E-Mailverteilung wurden am 29. April 5'185 E-Mails an den öffentlichen Sektor resp. am 2. Mai 19'163 Reminder-E-Mails an Universitäten und Hochschulen gesendet. Gelegentlich kam es zu

Rückfragen zur Studie oder dem Fragebogeninhalt, Interessenbekundungen an der Thematik sowie an der Umfrage teilzunehmen und diese im eigenen Netzwerk weiterzuleiten und Absagen. Auf diese Antworten wurde stets zeitnah und ausführlich geantwortet, wobei die Chance dieses persönlichen Kontakts genutzt wurde, um die Personen zum Weiterleiten des Umfragelink innerhalb ihres Netzwerks zu motivieren.

Generell wurde eine Samplegrösse von 200 Respondenten angestrebt, um geeignete statistische Auswertungsmethoden anzuwenden. Insgesamt wurden 253 Respondenten rekrutiert, wovon 189 Fälle nicht die zwei für die Befragung notwendigen Kriterien erfüllten und ausgeschlossen wurden. Die anhand zweier Filterfragen gemessener Kriterien mussten insofern erfüllt sein, dass die Respondenten bei Evaluationen die Rolle von Auftraggebern (und nicht von Evaluierenden (n = 134) oder Evaluationsdienstleistern oder ähnliches (n = 25)) einnehmen (n = 94) und bereits alleine oder im Team eine Evaluation beauftragt haben (n = 64). Da im Rahmen der Studie keine Bewilligung zur Kontaktaufnahme von staatlichen Behörden des Bundesstaates Washington D.C. in bezüglich Forschungsvorhaben vorlag, wurden alle betroffenen Fälle (n = 2) mit der IP-Adresse aus Washington D.C. vollständig gelöscht. Somit resultierten 62 gültige und für die Datenauswertung verwendbare Fälle. Darunter existieren jedoch Fälle mit tiefen Erledigungsraten, wobei 42 Prozent (n = 26) die gesamte Umfrage vervollständigten. Dazu fällt auf, dass die höchste Abbruchsrate (11%, n = 7) im ersten Viertel des Umfrageverlaufs liegt. Das finale Sample setzt sich aus insgesamt 61 Prozent (n = 19) weiblichen Respondenten zusammen, wobei das Durchschnittsalter bei 46.3 Jahren (SD = 12.8) liegt und über 80.8 Prozent (n = 28) der Respondenten über mindestens einen Masterabschluss verfügen (N = 31) (weitere Angaben siehe Anhang A. Variablenübersicht). Im nächsten Abschnitt wird der Fragebogen als Erhebungsinstrument vorgestellt.

3.2 Fragebogenentwicklung

In folgendem Abschnitt wird auf den Aufbau des Fragebogens (siehe Anhang B. Externer Anhang, a. Codebuch), auf Besonderheiten und Formalitäten in der Fragebogenentwicklung sowie den Pretest eingegangen.

Insgesamt erstreckt sich die Online-Befragung über 39 A4-Seiten und enthält 39 Fragen. Nach der Fragebogenentwicklung wurde der finalisierte Fragebogen durch die Autorin ins Englische übersetzt. Die Mehrheit der Fragen konstituiert sich aus Verhaltensfragen, Fragen zu Befragteneigenschaften, Einstellungsfragen und Überzeugungsfragen, wobei die letzte offene Frage Rückmeldungen zur Umfrage zulässt (Schnell et al., 2018, S. 297–

300). Die Online-Befragung weist mit den mehrheitlich geschlossen gestellten Fragen und den wenigen offenen Fragen, die zur Ergänzung möglicher fehlender Antwortkategorien dienten, einen hohen Standardisierungsgrad auf (Diekmann, 2013, S. 437).

Dem ersten sichtbaren Bildschirm, welcher die Online-Befragung kurz vorstellt und einleitet, kommt eine zentrale Bedeutung für die Motivation zur Teilnahme von potenziellen Befragten zu. Sowohl der Gegenstand der Befragung als auch die *ZHAW,* als durchführende Institution, wurden hervorgehoben (Schnell et al., 2018, S. 348). Neben der Befragungsanleitung und maximalen Befragungsdauer wurde zudem auf die Bewahrung der Anonymität und vertraulichen Behandlung der Daten hingewiesen. Zur Vertrauensbildung und in Hinblick auf einen professionellen Auftritt wurden Kontaktinformationen angegeben und das ZHAW-Logo den ganzen Fragebogen hindurch als Header verwendet. Nach der Einführung in den Fragebogen inklusive Filterfragen gliedert sich der Fragebogen in die sechs thematischen Blöcke der Evaluationspraxis, Zusammenarbeit, Unzufriedenheit mit Evaluationen, Rolle als Auftraggeber, Evaluationsstandards und der Soziodemographie. Bei der Fragenbogenkonstruktion wurde generell das Prinzip der Trichterung verfolgt, bei dem durch die Abfolge der Fragen in den für den Frageblock relevanten Kontext eingeführt wird (Ebd., 2018, S. 313–314). Anfangs dienen die beiden bereits erwähnten Filterfragen dazu, nur die für die Studie relevanten Respondenten zu selektionieren. Generell wurden bei der Verneinung einer Filterfrage die dazugehörigen Anschlussfragen übersprungen, wobei die Umfrage bei einer „Ja"- oder „Weiss nicht"-Antwort fortgesetzt wurde. Zum Schluss wird den Befragten für die Unterstützung gedankt, wobei sie erneut an die Wichtigkeit zusätzlicher Umfrageteilnahmen erinnert und dazu motiviert werden den angefügten Umfragelink innerhalb ihres Netzwerks an weitere Auftraggeber von Evaluationen weiterzuleiten.

Generell wurden sensiblere und eher schwierig zu beantwortende Fragen ans Ende eines Fragenkomplexes innerhalb eines Themenblocks gesetzt. Würden diese Fragen ans Fragebogenende gestellt werden, würde dies zwar frühzeitige Abbrüche vermindern, die Fragen jedoch aus dem jeweiligen Kontext des thematischen Blocks werfen, was vorliegend verhindert wurde (Ebd., 2018, S. 314). Für die Programmierung des Fragebogens mithilfe der Survey Software von Qualtrics wurden v.a. die beiden Fragetypen „Multiple Choice" und der Matrixtabelle verwendet (Qualtrics, 2019c). Letzterer wurde insb. für die Verhaltens- und Einstellungsfragen verwendet, die über Skalen mit mehrstufigen Itembatterien gemessen werden. Bei der Formulierung der Fragen und Antwortkategorien wurden gewisse Regeln befolgt. Da sich das Leseverhalten am Bildschirm oft durch das Scannen

von Text statt durch genaues Lesen charakterisiert, wurde generell sparsam mit Text umgegangen. Die Fragetexte wurden auf das nötigste reduziert und übersichtlich präsentiert (Diekmann, 2013, S. 529). Zudem wurde darauf geachtet, dass die Fragen einfache Worte und keine doppelten Negationen enthalten, konkret, neutral und nicht hypothetisch formuliert sind. Suggestivfragen wurden vermieden, sodass keine bestimmte Antwort provoziert wurde. Weiter wurde die Eindimensionalität der Fragen eingehalten, indem sich die Fragen jeweils auf einen Sachbehalt beziehen (Schnell et al., 2018, S. 306). Darüber hinaus wurde den Antwortvorgaben fast aller Fragen, mit Ausnahme der personalen und soziodemographischen Fragen, eine explizite „Weiss-nicht"-Kategorie hinzugefügt, um einerseits bei ausgewählten Fragen diese Zusatzkategorie als inhaltlich valider und interpretierbarer Wert zu betrachten. Andererseits würden Respondenten zu einer Antwort gezwungen, auch wenn sie tatsächlich die Frage nicht beantworten können, was zu Verzerrungen in den Ergebnissen führen würde (Ebd., 2018, S. 308). Zur Reduktion des Non-Attitude-Problems dienten die eingeschobenen Filterfragen der jeweiligen Themenblöcke dazu, dass den Befragten nicht irrelevante Fragen angezeigt wurden (Diekmann, 2013, S. 454). Befragtenmerkmale können in der Befragung weitere Fehlerquellen aufweisen, die aus den Problemen des Response-Sets resp. der Akquieszenz und der sozialen Erwünschtheit resultieren. Gegen Response-Set und Akquieszenz wurde insofern vorgegangen, dass die Items sowohl in positiver und negativer Richtung auf die Zieldimensionen gepolt wurden (Ebd., 2013, S. 453). Zur Reduktion des Effekts der sozialen Erwünschtheit wurde der Fragetext zur Messung der destruktiven Beeinflussungsart suggestiv formuliert, wodurch ein abweichendes Verhalten als normal deklariert und somit der Ort der sozialen Erwünschtheit auf der Skala gewissermassen verschoben wird. Neben der gewährleisteten Anonymität der Befragung, wurde eine Skala zur Messung der *sozialen Erwünschtheit* verwendet (siehe Anhang A. Variablenübersicht). Die Idee dieser Skalenkonstruktion basiert darauf, dass Personen in unterschiedlichem Ausmass für den Effekt der sozialen Erwünschtheit empfänglich sind (Ebd., 2013, S. 447–451). Um sowohl den allgemeinen Eindruck des Fragebogens wie die Verständlichkeit, Logik, Kohärenz, der zeitliche Aufwand als auch die Übersetzung und die Güte der Filterführung des Fragebogens zu prüfen und mögliche Schwierigkeiten und Unsicherheiten zu identifizieren, wurde ein Pretest durchgeführt (Schnell et al., 2018, S. 317; Sedlmeier & Renkewitz, 2018, S. 113). Nach der Pretest-Durchführung wurde der Fragebogen aufgrund der ausführlichen Rückmeldungen überarbeitet. Nachdem hiermit die Umsetzung und der inhalt-

liche Aufbau des Online-Fragebogens beschrieben wurden, werden im nächsten Abschnitt die für die Fragebogenentwicklung relevanten Operationalisierungen der Konstrukte und Skalenentwicklung geschildert.

3.3 Operationalisierung der Konstrukte und Skalenentwicklung

Im folgenden Abschnitt werden sowohl die für die Überprüfung der Hypothesen relevanten Konstrukte beschrieben und operationalisiert als auch auf Besonderheiten der Skalenentwicklung wie die Indexbildung eingegangen. Zur ganzheitlichen Übersicht aller Variablen, ihrer Operationalisierungen sowie Ausprägungen inklusive zusammenfassende Statistiken wird auf den Anhang verwiesen (siehe A. Variablenübersicht).

Der Zusammenhang des Beziehungsverhältnisses zwischen dem Auftraggeber und dem Evaluierenden und der Einflussnahme des Auftraggebers auf den Evaluationsprozess wird anhand des ersten Hypothesenblocks getestet. Bei den drei ersten Hypothesen soll der Zusammenhang zwischen der abhängige Variable der *destruktiven Beeinflussungsart*[3] und der jeweiligen unabhängigen Variablen des *Konfliktverhältnisses,* der *Unzufriedenheit* und der *Schwierigkeiten* untersucht werden. Bei der vierten Hypothese wird wiederum der Zusammenhang der unabhängigen Variable des *Konfliktverhältnisses* mit der abhängige Variable des *Anreizsystems* (*H4a*) oder genauer des *direkten Einflusses* (*H4b*) analysiert. Die Operationalisierung sowie Skalenentwicklung dieser Variablen des ersten Hypothesenblocks wird nachfolgend beschrieben:

Konfliktverhältnis: Das Konstrukt des durch die Auftraggeber wahrgenommenen Konfliktverhältnisses wird durch die ordinale Variable *Konfliktverhältnis* gemessen. Der Interessenkonflikt wird somit nicht als Konstante, sondern als Variable mit unterschiedlichen Werten operationalisiert (Waterman & Meier, 1998, S. 185). Die Variable wurde anhand einer 11-stufigen Skala operationalisiert, die misst, als wie konfliktgeprägt Auftraggeber das Verhältnis zu Evaluierenden in ihrer Arbeitstätigkeit generell wahrnehmen. Die Respondenten konnten ihre Antwort anhand einer Skala von 0 (= überhaupt nicht konfliktgeprägt) bis 10 (= äusserst konfliktgeprägt) einordnen. Neben dieser Stärkemessung wurde ergänzend ein additiver Index zur *Konflikthäufigkeit* gebildet, der auf der dichotomen Skala zur Messung der Konfliktgründe basiert und in sieben Kategorien eingeteilt wurde, die sich von der Kategorie „nie“ (= 0) bis zur Kategorie „äusserst häufig“ (= 6) erstrecken. Der Homogenitätsgrad der Skala *Konflikthäufigkeit* liegt mit einem

[3] Die für die Hypothesen relevanten Variablen werden in vorliegender Arbeit jeweils kursiv dargestellt.

Cronbach's Alpha von .719 in einem guten Bereich und könnte durch den Ausschluss des Items „unterschiedliche Ansichten bezüglich des Auftrags und dessen Bedingungen" verbessert werden (Cronbach's Alpha: .724). Die Variable *Unterstellung* misst eine gewisse Form des Konfliktverhältnisses und kann ebenso in die Analysen integriert werden.

Unzufriedenheit: Das Konstrukt der Unzufriedenheitsstärke aufseiten der Auftraggeber wird durch die unabhängige ordinale Variable *Unzufriedenheit* gemessen. Für die Operationalisierung des Konstrukts wurde die Skala der Vorbildstudie erweitert, anhand welcher die Hauptgründe für die Unzufriedenheit mit der in Auftrag gegebenen Evaluation gemessen wurden (Pleger & Hadorn, 2018). Die Erweiterung besteht darin, dass nicht nur die Gründe, sondern vielmehr das Ausmass der Unzufriedenheit hinsichtlich sieben unterschiedlicher Aspekte einer Evaluation (Items) gemessen wird. Unter anderem wird die Einschätzung der Unzufriedenheit bezüglich der Evaluationsresultate und -schlussfolgerungen, der Evaluations- und Methodenkompetenz der Evaluierenden und der Evaluationsqualität abgefragt. Die Skala bezieht sich auf die letzte beauftrage Evaluation, mit welcher die Auftraggeber unzufrieden waren, wobei die Befragten ihre Antwort zwischen den Werten 0 (= überhaupt nicht zufrieden) und 10 (= voll und ganz zufrieden) einordnen konnten. Basierend auf der Skala wurde ein additiver Index gebildet, wofür alle Items in diejenige Richtung umgepolt werden mussten, sodass hohe Werte mit einer hohen Unzufriedenheit korrespondieren. Der Werterange des Index beläuft sich von 0 bis 70, wobei die Werte wiederum den Kategorien von 0 (= sehr zufrieden) bis 4 (= sehr unzufrieden) zugeteilt wurden. Der Homogenitätsgrad der Skala ist mit einem Cronbach's Alpha Wert von .853 sehr hoch und könnte einzig durch den Ausschluss des Items „Einhalten des Zeitplans" verbessert werden (Cronbach's Alpha: .916).

Schwierigkeiten: Das Konstrukt der Schwierigkeiten in der Zusammenarbeit mit Evaluierenden wird durch die ordinale Variable *Schwierigkeiten Häufigkeiten* gemessen. Die Operationalisierung dieses Konstrukts basiert auf den Resultaten zu den wahrgenommenen Hauptschwierigkeiten bei der Zusammenarbeit mit Evaluierenden aus der Studie von Pleger und Hadorn (2018), die in fünf Kategorien unterteilt wurden. Daraus abgeleitet wurde eine dichotome Skala mit sieben eindimensionalen Items gebildet, welche Schwierigkeiten –wie bspw. ein fehlendes Verständnis für die zu evaluierende Organisation, ein fehlendes gegenseitigen Verständnis, zu wenigen Ressourcen, mangelhafte Fachkompetenzen oder eine fehlende Motivation – beinhalten. Zur Messung der Variable wurde da-

raus ein additiver Index gebildet, der die Häufigkeiten der unterschiedlichen Schwierigkeiten zusammenfasst und diese wiederum in sieben Kategorien einteilt. Der Cronbach's Alpha Wert der Skala ist mit einem Wert von .477 sehr tief, wobei das Item bezüglich des Einflusses persönlicher Faktoren ausgeschlossen wird und sich der Homogenitätsgrad der Skala auf einen Wert von .538 verbessert. Da das Konstrukt der Schwierigkeiten sehr breit gefasst ist, wird dieser Wert für die vorliegende Studie so akzeptiert (Krüger, Borgmann, & Antonik, 2012, S. 48).

Destruktive Beeinflussungsart: Die destruktive Einflussnahme der Auftraggeber auf den Evaluationsprozess resp. auf die Evaluierenden wird als Konstrukt anhand der ordinalen Variable *destruktive Beeinflussungsart* gemessen. Für die Operationalisierung des Konstrukts wurde die 4-stufige Skala der Schweizer Studie adaptiert, anhand welcher das Ausmass der destruktiven Beeinflussung – hauptsächlich *Distortion* – mit acht Items gemessen wird (Pleger & Hadorn, 2018). Adaptiert wurde die Skala insofern, dass sie anhand des BUSD-Modells operationalisiert wurde, wobei beide Beeinflussungsformen *Distortion* und *Undermining* inkludiert und jeweils anhand vier Items gemessen wurden (siehe Kapitel 2.3.1). Die dazugehörige Frage lautet: „Es ist völlig normal, dass Auftraggeber in irgendeiner Weise in den Evaluationsprozess intervenieren. Welche der nachfolgenden Änderungsvorschläge an einer Evaluation haben Sie in Ihrer bisherigen Karriere als AuftraggeberIn geäussert?". Die Antworten konnten anhand von vier Häufigkeitskategorien[4] eingeordnet werden. Basierend auf der Skala wurde ein additiver Index mit dem Werterange von 0 bis 24 gebildet, wobei diese wiederum den vier Häufigkeitskategorien zugeteilt wurden. Als kontrollierende Ergänzung wurde eine zweite ordinale Variable *destruktive Beeinflussungsart Allgemein* erstellt, die das Konstrukt der destruktiven Einflussnahme relativ genereller misst und als Vergleich zur Variable der *destruktiven Beeinflussungsart* in die Berechnungen integriert wird. Dabei wurden für die beiden Beeinflussungsformen *Distortion* und *Undermining* charakteristische Merkmale gemäss dem BUSD-Modell aufgegriffen und anhand zweier Items hinsichlich eines möglichen Vorkommnisses operationalisiert: Das Item des direkten, negativen Beeinflussungstyps *Distortion* beschreibt die mögliche Aufforderung, dass gewisse Evaluationsresultate gekürzt werden, der indirekte, implizite Beeinflussungstyp *Undermining,* dass es aufgrund der Änderungsvorschläge zu einer Verzögerung des Evaluationsprozesses kommen kann. Die

[4] Sowohl die Skalen als auch die Indizes der Variablen *destruktive Beeinflussungsart*, *destruktive Beeinflussungsart Allgemein, Anreizsystem* und *direkter Einfluss* bestehen aus den folgenden Ausprägungen: Nein, habe ich noch nie gemacht (0), Ja, habe ich einmal gemacht (1), Ja, habe ich mehr als einmal gemacht (2), Ja, habe ich oft gemacht (3)

Antworten konnten mittels der vier erwähnten Häufigkeitskategorien angegeben werden. Daraus wurde ein additiver Index mit dem Werterange von 0 bis 6 gebildet und erneut den Häufigkeitskategorien zugeordnet. Der Homogenitätsgrad der Skala *destruktive Beeinflussungsart* liegt mit einem Cronbach's Alpha von .623 in einem tiefen Bereich und kann durch den Ausschluss des Items „Ich habe wichtige Informationen gegenüber dem Evaluator zurückgehalten" auf den Alphawert von .635 erhöht werden. Angesichts des breiten Konstrukts wird dieser Wert vorliegend akzeptiert. Für die Skala *destruktive Beeinflussungsart Allgemein* ist Cronbach's Alpha mit einem Wert von .523 sehr tief, wird aber angesichts der kurzen Skala akzeptiert.

Anreizsystem: Das Konstrukt des Anreizsystems, mithilfe dessen ein Auftraggeber versucht dem Interessenkonflikt mit einem Evaluierenden entgegenzuwirken und den Evaluierenden dadurch zu beeinflussen (Roiger, 2007, S. 1), wird anhand der ordinalen Variable *Anreizsystem* gemessen. Für die Operationalisierung des Konstrukts wurde die 4-stufige Skala von Pleger und Hadorn (2018) insofern adaptiert, dass sowohl die direkte, explizite als auch die indirekte, implizite Beeinflussungsart durch jeweils drei Items gemessen wurde. Die Items wurden anhand des BUSD-Modells (siehe Kapitel 2.3.1) operationalisiert und somit theoretisch abgeleitet. Der direkte Einfluss wurde über die drei Items gemessen, dass dem Auftraggeber erstens klare Anreize für Veränderungen der Resultate gegeben werden und zweitens der Evaluationsauftrag entzogen werden könnte. Das dritte Item wurde über die Aufforderung zur Änderung der Resultate unter Androhung negativer Konsequenzen gemessen. Die drei Items des indirekten Einflusses beinhalteten die Androhung das Honorar nicht zu bezahlen, den Bericht nicht zu veröffentlichen und den Evaluierenden nicht für zukünftige Ausschreibungen zu berücksichtigen. Die Antworten konnten anhand der bereits erläuterten Häufigkeitskategorien angegeben werden. Basierend auf dieser Skala wurde für die Variable *Anreizsystem* ein additiver Index mit einem Werterange von 0 bis 18 gebildet und denselben Häufigkeitskategorien zugeteilt. Der Homogenitätsgrad der Skala liegt mit einem Cronbach's Alpha von .714 in einem guten Bereich und kann durch den Ausschluss des Items „dem Evaluator in Aussicht stellen, das Honorar nicht zu zahlen" auf den Wert .743 verbessert werden.

Direkter Einfluss: Das Konstrukt des direkten, expliziten Einflusses wird anhand der ordinalen Variable *direkter Einfluss* gemessen. Als inhärenter Bestandteil des Anreizsystems wird bei der Messung dieser Variable auf die Skala des Anreizsystems zurückgegriffen. Basierend auf den drei Items, welche die Dimensionen des direkten Einflusses

abdecken, wird erneut ein additiver Index mit dem Werterange von 0 bis 9 erstellt und den gleichen Häufigkeitskategorien zugeteilt. Angesichts der kurzen Skala wird der zwar eher tiefe Cronbach's Alpha Wert von .678 akzeptiert, wobei ein Itemausschluss nicht sinnvoll ist.

Im zweiten Hypothesenblock wird der Zusammenhang der Eigenschaften der Auftraggeber mit der Erwartung an die Unabhängigkeit von Evaluationen sowie der konstruktiven Beeinflussungsart der Auftraggeber getestet. Die zusammenhängenden Effekte der Eigenschaften der Auftraggeber – wie die *Vertrautheit* mit Standards und die *Berufserfahrung* – als unabhängige Variablen werden mittels der sechsten und siebten Hypothesen (*H6-H7*) und der *konstruktiven Beeinflussungsart* als abhängige Variable überprüft. Bei der fünften Hypothese (*H5*) resultiert die *Vertrautheit* erneut als unabhängige Variable, wobei deren Zusammenhang mit der abhängigen Variable der *Erwartung an Unabhängigkeit* analysiert wird. Die Operationalisierung sowie Skalenentwicklung dieser Variablen des zweiten Hypothesenblocks wird nachfolgend beschrieben:

Vertrautheit: Die Vertrautheit mit den nationalen „Program Evaluation Standards" wird als Konstrukt über die ordinale Variable *Vertrautheit* gemessen. Dabei wird das Konstrukt über die von der Schweizer Studie übernommenen 4-stufigen Skala operationalisiert und auf den US-Evaluationskontext übertragen (Pleger & Hadorn, 2018).

Erwartung an Unabhängigkeit: Das Konstrukt der Erwartung an die Unabhängigkeit von Evaluationen wird durch die ordinale Variable *Erwartung an Unabhängigkeit* gemessen. Die Operationalisierung dieses Konstrukts basiert auf der 6-stufigen Skala aus der Studie von Pleger und Hadorn (2018) sowie Stockmann et al. (2011), welche die Wahrnehmung ethischer Standards für Evaluierende misst. Genauer fragen die Items der Skala was nach der Auffassung des Auftraggebers einen guten Evaluierenden ausmacht, wobei die Skala an wichtigen Bestandteilen der Guiding Principles der AEA anknüpft (AEA, 2011). Die Befragten konnten ihre Antwort in Form einer Zustimmung auf einer Skala von 0 bis 5 einordnen. Von der ursprünglichen Skala wurden zwei Items eliminiert, die darin bestanden, dass ein guter Evaluierender einen Antrag aufgrund von mangelnden Ressourcen resp. einer nicht sachbezogenen Fragestellung ablehnt. Somit umfasst die neue Skala insgesamt acht Items, wobei das erste, dritte, fünfte und sechste Item in ihrer Richtung umgedreht wurden. Für die Bildung des additiven Index werden diese Items entsprechend umgepolt. Der Werterange des Index reicht von 0 bis 40 und wird erneut

sechs Kategorien[5] zugeteilt, welche nicht die Zustimmung, sondern die Intensität der Erwartung an die Unabhängigkeit von Evaluationen widerspiegelt. Der Wert von Cronbach's Alpha von .682 weist auf einen tiefen Homogenitätsgrad der Skala hin und kann durch keinen Itemausschluss verbessert werden. Angesichts des breiten Konstrukts wird der Wert vorliegend akzeptiert.

***Berufserfahrung*:** Das Konstrukt der Berufserfahrung von Auftraggebern wird anhand der Variable *Berufserfahrung* gemessen, die sich als additiver Index aus den Variablen *Evaluationsjahre* und *Evaluationsanzahl* zusammensetzt. Die Variable *Evaluationsjahre* wurde anhand einer 5-stufigen Skala und die Variable *Evaluationsanzahl* mittels einer 3-stufigen Skala gemessen. Die Skalen wurden von Pleger et al. (2016) adaptiert. Aus beiden Variablen wurde ein additiver Index mit dem Werterange von 0 bis 6 gebildet, wofür sieben neue Kategorien[6] gebildet wusrden, um das Ausmass der Erfahrung darzustellen. Der Homogenitätsgrad der Skala liegt mit einem Cronbach's Alpha von .550 sehr tief. Aufgrund der kurzen Skala können keine Items ausgeschlossen werden und der Wert wird entsprechend akzeptiert.

Konstruktive Beeinflussungsart: Das Konstrukt der konstruktiven Einflussnahme der Auftraggeber auf den Evaluationsprozess resp. auf die Evaluierenden wird mittels der ordinalen Variable *konstruktive Beeeinflussunsgart* gemesssen. Abgeleitet vom BUSD-Modell wird für die Operationalisierung des Konstrukts eine neue 4-stufige Skala erstellt, die das Ausmass der konstruktiven Beeinflussung und der ihr inhärenten Beeinflussungsformen des *Supports* und *Betterments* misst (siehe Kapitel 2.3.1). Beide Dimensionen werden jeweils anhand von vier Items gemessen. Die dazugehörige Frage lautet: „Wenn Sie an die Zusammenarbeit mit den bisherigen Evaluierenden im Rahmen eines Evaluationsauftrages denken, welche der folgenden Handlungen haben Sie in Ihrer bisherigen Karriere als AuftraggeberIn ausgeführt?". Analog zur Variable *destruktive Beeinflussungsart* konnten die Antworten anhand derselben vier Häufigkeitskategorien eingeordnet werden. Basierend auf der Skala wurde ein additiver Index mit dem Werterange von 0 bis 24 gebildet, wobei diese wiederum den vier Häufigkeitskategorien zugeteilt wurden. Ebenfalls wurde ergänzend eine zweite ordinale Variable *konstruktive Beeinflussungsart Allgemein* erstellt, die das Konstrukt der konstruktiven Einflussnahme generell misst und

[5] Die Kategorien des Index der Variable *Erwartung an Unabhängigkeit* besitzen folgende Ausprägungen: Gar nicht (0), Tief (1), Eher tief (2), Eher hoch (3), Hoch (4), Sehr hoch (5).
[6] Die Kategorien des Index der Variable *Berufserfahrung* besitzen folgende Ausprägungen: Äusserst unerfahren (0), Sehr unerfahren (1), Eher unerfahren (2), Weder noch (3), Eher erfahren (4), Sehr erfahren (5), Äusserst erfahren (6).

als Vergleich zur Variable der *konstruktiven Beeinflussungsart* in die Berechnungen miteinfliesst. Theoretisch abgeleitet vom BUSD-Modell wurde für die beiden Beeinflussungsformen *Betterment* und *Support* jeweils ein Item gebildet: Das Item des direkten, positiven Beeinflussungstyps *Betterment* beschreibt die proaktive Illustration von Verbesserungspotenzial, der indirekte, implizite Beeinflussungstyp *Support* beschreibt das generelle Engagement aufseiten der Auftraggeber die Evaluationsqualität zu optimieren. Analog zur Variable der *destruktiven Beeinflussungsart Allgemein* konnten die Antworten anhand derselben vier Häufigkeitskategorien angegeben werden. Daraus wurde ein additiver Index mit dem Werterange von 0 bis 6 gebildet und erneut den Häufigkeitskategorien zugeordnet. Der Homogenitätsgrad der Skala der *konstruktiven Beeinflussungsart* liegt mit einem Cronbach's Alpha von .842 in einem sehr guten Bereich. Derjenige der *konstruktiven Beeinflussungsart Allgemein* ist mit einem Alphawert von .673 eher tief, angesichts der kurzen Skala jedoch in einem akzeptablen Bereich.

Ergänzende Variablen: Als mögliche Kontrollvariablen fungieren die soziodemographischen Variablen des *Alters*, der *Bildung* und des *Geschlechts,* sowie die *soziale Erwünschtheit* und die *Beeinflussungsintention.* Die beiden Variablen der *konstruktiven* sowie *destruktiven Beeinflussungsart Allgemein* fungieren als Vergleich zu den Variablen der *konstruktiven* und *destruktiven Beeinflussungsart*. Der Miteinbezug weiterer Variablen führt zu ergänzenden Erkenntnissen (siehe Anhang A. Variablenübersicht). Nachdem die Operationalisierung der Konstrukte und Skalenentwicklung vorgestellt wurden, wird im nächsten Abschnitt die Vorgehensweise bei der Datenauswertung beschrieben.

3.4 Datenauswertungsmethoden

Das Ziel dieser explorativen Studie liegt einerseits in der Überprüfung der aufgestellten Hypothesen und andererseits im Vergleich sowie der Validitätsprüfung der Befunde der Schweizerstudie (Pleger & Hadorn, 2018). Die dafür verwendeten Datenauswertungsmethoden werden nachfolgend erörtert.

Der Datensatz besteht v.a. aus ordinalen unabhängigen und abhängigen Variablen, wobei sich die Kontrollvariablen und ergänzenden Variablen zusätzlich auf das nominale und in Ausnahmefällen das metrische Skalenniveau stützen. Aufgrund des ordinalen Skalenniveaus der zu untersuchenden Variablen werden für die Datenauswertung mittels SPSS nicht-parametrische exakte Tests angewendet, die sich für nicht-normalverteilten Daten

eignen (Bühl, 2014, S. 408). Um die Normalverteilung der Daten zu testen, wird der Kolmogorov-Smirnov-Test verwendet (Ebd., 2014, S. 380). Für die Hypothesenüberprüfung eignet sich die Berechnung von Korrelationen und Assoziationen. Für Variablen mit ordinalem Skalenniveau bietet sich für die Berechnung von Korrelationen der Rangkorrelationskoeffizient von Spearman oder Kendalls Tau an (Dormann, 2017, S. 92). Der Kendall-Tau-b-Koeffizient ist bei Vorhandensein von Ausreissern weniger empfindlich gegenüber Verzerrungen und bei nicht-normalverteilten Daten und kleinen Stichproben dem Spearman-Rho vorzuziehen (Hochschule Luzern, 2019). Ausserdem werden beim spearmanschen Rangkorrelationskoeffizienten die Rangplätze als intervallskalierte und nicht ordinalskalierte Werte aufgefasst (Benninghaus, 2007, S. 184). Da für die vorliegende Studie nicht die Voraussetzung geschaffen wird, ordinale Daten als intervallskaliert zu definieren, wird das für ordinalskalierte Daten häufig angewendete Zusammenhangsmass Kendalls Tau berechnet (Sedlmeier & Renkewitz, 2018, S. 239). Die ordinalen oder nominal dichotomisierten Variablen, lassen sich ebenso in einer Kontingenztabelle zusammenfassen, wobei der X^2-Test und Fishers Exakter Test zum Test auf Assoziation angewendet werden kann (Dormann, 2017, S. 95). Wenn die Stichprobengrösse kleiner oder gleich 20 beträgt oder erwartete Zellhäufigkeiten von kleiner als fünf vorliegen, wird der exakte Test nach Fisher empfohlen (Universität Zürich, 2019). Da diese Voraussetzung bei beinahe allen untersuchten Variablen zutrifft, wird hauptsächlich Fishers Exakter Test gerechnet. Der X^2-Test scheint bei geringen erwarteten Häufigkeiten robust zu sein und zu annähernd korrekten p-Werten zu führen, wobei sich die Wahrscheinlichkeit für einen α-Fehler auch bei erwarteten Häufigkeiten von kleiner als 1 nicht nennenswert ändert (Sedlmeier & Renkewitz, 2018, S. 569). Bei kleinen Fallzahlen lassen sich bei SPSS zudem die exakten p-Werte bestimmen (Bühl, 2014, S. 408). Zur Beurteilung der gefundenen Effektgrössen werden vorliegend die Richtwerte nach Cohen (1988)[7] verwendet. Neben diesen Berechnungen informieren deskriptive Statistiken über die Charakteristika der jeweiligen Variablen (siehe Anhang A. Variablenübersicht). Abschliessend werden die Befunde zu den US-Auftraggebern von Evaluationen mit denjenigen aus der Schweiz in Hinblick auf individuelle Eigenschaften und das Beziehungsverhältnis zwischen Auftraggebern und Evaluierenden verglichen und Gemeinsamkeiten sowie Unterschiede beschrieben. Das nächste Kapitel stellt die gerechneten Zusammenhänge zur Hypothesenüberprüfung, die ergänzenden Befunde sowie die Resultate des Vergleichs zwischen Auftraggebern der USA und der Schweiz vor.

[7] Richtwerte von Cohen (1988): kleiner Effekt = 0.1, mittlerer Effekt = 0.3, grosser Effekt = 0.5

4 Resultate

In diesem Kapitel werden die Hypothesen anhand der durch die Online-Befragung erhobenen Daten dahingehend überprüft, ob Hinweise für eine vorläufige Bestätigung oder ein Nichtzutreffen der Hypothesen vorliegen. Dazu werden anhand verschiedener Variablen oder Items Indikatoren dafür oder dagegen gesammelt. Der erste Hypothesenblock wird im ersten Abschnitt, der Zweite im zweiten Abschnitt besprochen. Für die Überprüfung der Hypothesen wurde das Statistikprogramm SPSS verwendet. Darüber hinaus werden im dritten Abschnitt ergänzende Resultate zur Beantwortung der Forschungsfragen beschrieben und im vierten Abschnitt Vergleiche zur Schweizer Studie von Pleger und Hadorn (2018) gezogen. Abschliessend werden die Gütekriterien der Objektivität, Reliabilität und Validität auf die vorliegende Studie angewendet.

Vor den eigentlichen Berechnungen wurden die unabhängigen und abhängigen Variablen anhand des Kolmogorov-Smirnov-Tests auf Normalverteilung überprüft (Bühl, 2014, S. 380). Mit Ausnahme der Variable *Konflikthäufigkeit* (p = .200), bei welcher die Werte hinreichend normalverteilt sind, wurden für alle Variablen eine statistisch signifikante Abweichung von der Normalverteilung gefunden ($p < .05$). Neben den kleinen Stichprobengrössen der einzelnen Variablen verdeutlichen die nicht-normalverteilten Daten erneut, dass die Verwendung von nicht-parametrischen Tests sinnvoll ist.

4.1 Resultate zum Beziehungsverhältnis zwischen Auftraggebern und Evaluierenden

Der Zusammenhang des Beziehungsverhältnisses zwischen Auftraggebern und Evaluierenden und der Beeinflussung des Evaluationsprozesses seitens der Auftraggeber wird anhand des ersten Hypothesenblocks getestet. Genauer testen die drei ersten Hypothesen wie die abhängige Variable der *destruktiven Beeinflussungsart* mit den jeweiligen unabhängigen Variablen des *Konfliktverhältnisses,* der *Unzufriedenheit* und der *Schwierigkeiten* zusammenhängt.

4.1.1 Destruktive Beeinflussungsart

Bei der minim linksschief-verteilten (Schiefe: -.007) Variable *destruktive Beeinflussungsart* haben 85 Prozent[8] der Befragten (n = 22) mindestens schon einmal destruktiven Einfluss auf Evaluierende ausgeübt (siehe Tabelle 2). Darunter haben 17 Personen (65%)

[8] Die Prozentwerte werden der Einfachheit halber im Fliesstext gerundet dokumentiert.

einmal Änderungsvorschläge geäussert, die der destruktiven Beeinflussungsart entsprechen, wobei dies fünf Personen (19%) mehr als einmal gemacht haben. Lediglich vier Respondenten (15%) haben keinen destruktiven Einfluss ausgeübt ($N = 26$).

Tabelle 2: Übersicht der Resultate zur destruktiven Beeinflussungsart

	0	1	2	3	*n*
Positivere Darstellung der Evaluationsresultate	74.1 (20)	7.4 (2)	18.5 (5)	-	27
Umformulierung wertender Sätze	37.0 (10)	3.7 (1)	29.6 (8)	29.6 (8)	27
Mitteilung erwarteter Ergebnisse	74.1 (20)	11.1 (3)	11.1 (3)	3.7 (1)	27
Weglassen gewisser Teile	70.4 (19)	18.5 (5)	7.4 (2)	3.7 (1)	27
Zurückhalten wichtiger Informationen	96.3 (26)	-	-	3.7 (1)	27
Neuverhandlung der Gebühren	74.1 (20)	3.7 (1)	11.1 (3)	11.1 (3)	27
Wiederholtes Nachfragen für Modifikation	44.4 (12)	18.5 (5)	29.6 (8)	7.4 (2)	27
Hinweis auf Vorgehensweise anderer	61.5 (16)	15.4 (4)	19.2 (5)	3.8 (1)	26
Destruktive Beeinflussungsart	15.4 (4)	65.4 (17)	19.2 (5)	-	26
Destruktive Beeinflussungsart Allgemein	48.1 (13)	33.3 (9)	14.8 (4)	3.7 (1)	27
Undermining	63.3 (19)	6.7 (2)	23.3 (7)	6.7 (2)	30
Distortion	65.5 (19)	20.7 (6)	10.3 (3)	3.4 (1)	29

Anmerkungen: Skala von 0 (= Nein, das habe ich nie gemacht), 1 (= Ja, das habe ich einmal gemacht), 2 (= Ja, das habe ich mehr als einmal gemacht), 3 (= Ja, das habe ich oft gemacht); *n* = Gesamtzahl aller valider Fälle. Die Zahlen repräsentieren Prozentwerte, in Klammern wurden die jeweiligen Häufigkeiten angegeben.

(Quelle: eigene Darstellung)

Die Forderung nach der Umformulierung einzelner wertender Sätze im Evaluationsbericht, scheint ein relativ weitverbreitetes Phänomen innerhalb der Auftraggeber zu sein. Die Mehrheit der Respondenten (63%, $n = 17$) haben dies mindestens einmal gemacht. Darunter haben jeweils 30 Prozent ($n = 8$) dies schon mehr als einmal resp. schon oft gemacht. Auch das mehrmalige Nachfragen, ob der Fragebogen oder Report modifiziert werden kann, scheint unter den Auftraggebern relativ üblich zu sein, wobei dies die Mehrheit (56%, $n = 15$) bereits mindestens einmal gemacht hat ($N = 27$). Der Hinweis darauf wie andere Evaluierende vorgegangen wären, wurde von zehn Personen (38%) mindestens einmal gemacht. Darunter haben dies vier Personen (15%) einmal, fünf Personen (19%) mehr als einmal und eine Person (4%) oft gemacht ($N = 26$). Das schon mindestens einmal gefordert wurde, gewisse Teile wegzulassen, haben insgesamt acht Befragte (30%) bejaht resp. hat die Mehrheit (70%) verneint ($N = 27$). Ungefähr ein Viertel der Befragten (26%, $n = 7$) haben schon mindestens einmal vorgeschlagen, dass die Evaluationsresultate positiver dargestellt werden sollen, wobei fünf Personen (19%) angaben, dies mehr als einmal gemacht zu haben und 74 Prozent ($n = 20$) dies noch nie gemacht haben ($N = 27$). Ebenso haben 20 Personen (74%) angegeben, die Gebühren der Evaluation noch nie neuverhandelt zu haben, was bei wiederum 26 Prozent ($n = 7$) mindestens einmal vorgekommen ist ($N = 27$). Dass den Evaluierenden im Vorhinein der Evaluation mitgeteilt wird, welche Ergebnisse erwartet werden, ist weniger verbreitet, wurde aber

ebenfalls von gut einem Viertel (*n* = 7) mindestens einmal gemacht. Eine Person (4%) gab sogar an, dies oft gemacht zu haben (*N* = 27). Nur eine Person hat angegeben schon oft wichtige Informationen gegenüber Evaluierenden zurückgehalten zu haben, wobei 96 Prozent (*n* = 26) dies verneinten (*N* = 27). Die Antworten zu der offenen Frage nach anderen Interventionen im Evaluationsprozess enthielten vordergründig Erklärungen zu den gegebenen Antworten auf die geschlossenen Fragen. Jemand ergänzte die eigene Antwort mit der Aussage, dass „Änderungsempfehlungen gemacht werden, um die Ergebnisse für ein breites Publikum besser verständlich zu machen und nicht, um Ergebnisse oder Konsequenzen von Befunden zu verändern". Eine weitere Person führte aus, dass ihre Evaluierenden nutzenfokussiert und nicht traditionelle „impact evaluators" seien und es daher durchaus angemessen sei, einen Fragebogen so zu modifizieren, dass er sich mit der eigenen Arbeit, dem Kontext und seiner Umwelt deckt. Zudem wurde ergänzt, dass Forderungen nach Änderungen am Fragebogen geäussert wurden, um sachdienlichere Informationen zu erhalten. Eine andere Person gab an, den Evaluierenden aufgefordert zu haben, den Evaluationsbericht vertragsgemäss zu vervollständigen. Gemäss der rechtsschiefen Variable *destruktive Beeinflussungsart Allgemein* (Schiefe: .943) haben 48 Prozent der Auftraggeber (*n* = 13) generell noch nie einen destruktiven Einfluss auf Evaluierende ausgeübt, wobei 52 Prozent der Auftraggeber (*n* = 14) einer generell destruktiven Beeinflussungsart zugewiesen werden können. Darunter haben im Allgemeinen jeweils neun Personen (33%) einmal, vier Personen (15%) mehr als einmal und eine Person (4%) oft auf destruktive Art und Weise den Evaluationsprozess beeinflusst (*N* = 27). Für die Variable *Undermining* als indirekter, destruktiver Beeinflussungstyp zeigt sich, dass für 37 Prozent (*n* = 11) der Auftraggeber mindestens einmal vorgekommen ist, dass sich der Evaluationsprozess aufgrund der Änderungsvorschläge etwas verzögert hat. Bei zwei Personen (7%) ist dies einmal, bei sieben Personen (23%) mehr als einmal und bei zwei Personen (7%) oft vorgekommen (*N* = 30). Beim direkten, destruktiven Beeinflussungstyp *Distortion* zeigt sich ein ähnliches, leicht abgeschwächtes Bild, wobei 35 Prozent der Auftraggeber (*n* = 10) mindestens einmal Evaluierende aufgefordert haben gewisse Evaluationsresultate zu kürzen. Darunter haben dies jeweils sechs Personen (21%) einmal, drei Personen (10%) mehr als einmal und eine Person (3%) oft gemacht (*N* = 29*)*. Die Variable *Beeinflussungsintention* ist zwar rechtsschief verteilt (Schiefe: .456), doch bereits alle Werte mit Ausnahme von Null – was einer konstruktiven Beeinflussungsart entspricht – implizieren eine destruktive Beeinflussungsart mit einer jeweiligen Intensität.

Die *Beeinflussungsintention* verdeutlicht, dass alle Respondenten ($N = 26$) eine destruktive Beeinflussungsform mit unterschiedlichen Intensitäten ausüben. Dabei äussert sich die Beeinflussungsintention bei 69 Prozent der Auftraggeber ($n = 18$) als wenig destruktiv, für 27 Prozent ($n = 7$) als mittelmässig destruktiv und für eine Person (4%) als sehr destruktiv.

4.1.2 Konfliktverhältnis

Die Variable *Konfliktverhältnis* ist rechtsschief verteilt (Schiefe: .422) und zeigt, dass 75 Prozent ($n = 27$) der Auftraggeber ihr Verhältnis zu Evaluierenden in ihrer Arbeitstätigkeit generell unter dem Wert 4 (von insgesamt 10) einordnen und der Wert 6 als Antwort nicht überschritten wurde. Demzufolge wird das Beziehungsverhältnis verhältnismässig als eher weniger konfliktgeprägt wahrgenommen. Die Filterfrage, ob die Auftraggeber jemals mit einem Evaluierenden aufgrund seiner Arbeit oder Arbeitsweise einen *Konflikt* hatten, wurde mit je 48.65 Prozent verneint resp. bejaht ($n = 18$) und einmal mit „Weiss nicht" (2.7%) beantwortet ($N = 37$). Bei der Variable *Konflikthäufigkeit* wurden aufgrund der Filterfrage nur Antworten von Auftraggebern erfasst, die in der Vergangenheit bereits einen Konflikt mit einem Evaluierenden hatten. Die Konflikthäufigkeiten sind tendenziell gleichmässig normalverteilt, wonach Konflikte von nie (5%, $n = 3$; $N = 17$) bis äusserst häufig (2%, $n = 1$) stattfinden.

Zur Überprüfung des angenommenen positiven Zusammenhangs zwischen dem konfliktgeprägten Verhältnis[9] und der *destruktiven Beeinflussungsart* von H1 wurden Rangkorrelationskoeffizienten von Kendall-Tau-b und Chi-Quadrat-Tests berechnet (siehe Tabelle 3). Der Korrelationskoeffizient von Kendall-Tau-b liegt bei $r_{tb} = .015$ ($p = .931$, $n = 25$). Gemäss dem exakten Test nach Fisher und dem Tau-c-Koeffizienten als symmetrisches Mass existiert kein statistisch signifikanter Zusammenhang zwischen dem *Konfliktverhältnis* und der *destruktiven Beeinflussungsart* ($p = .572$; $\tau_c = .014$, $p = .943$).

Wird anstatt der *destruktiven Beeinflussungsart* die Variable *destruktive Beeinflussungsart Allgemein* in die Berechnungen integriert, zeigt sich für das *Konfliktverhältnis* ebenso ein statistisch nicht signifikanter, jedoch negativer Zusammenhang. Der Rangkorrelationskoeffizient von Kendall-Tau-b beträgt $r_{tb} = -.029$ ($p = .863$, $n = 26$). Gemäss dem exakten Test nach Fisher und dem Tau-c-Koeffizienten existiert kein statistisch signifikanter Zusammenhang zwischen dem *Konfliktverhältnis* und der *destruktiven Beeinflussungsart Allgemein* ($p = .403$; $\tau_c = -.028$, $p = .875$).

[9] Zur Berechnung des konfliktgeprägten Verhältnisses werden die beiden Variablen *Konfliktverhältnis* und *Konflikthäufigkeit* nacheinander getestet. Dies gilt für H1 sowie H4.

Tabelle 3: Übersicht der Resultate zur Überprüfung von H1

	Rangkorrelation			Chi-Quadrat-Tests		
	Kendall-Tau-b			Exakter Test nach Fisher	Kendall-Tau-c	
	Korrelations-koeffizient r_b	Signifikanz (2-seitig)	*N*	Exakte Signifikanz (2-seitig)	Wert τ_c	Exakte Signifikanz
Konfliktverhältnis x *destruktive Beeinflussungsart*	.015	.931	25	.572	.014	.943
Konfliktverhältnis x *destruktive Beeinflussungsart Allgemein*	-.029	.863	26	.403	-.028	.875
Konflikthäufigkeit x *destruktive Beeinflussungsart*	.372	.159	12	.945	.313	.183
Konflikthäufigkeit x *destruktive Beeinflussungsart Allgemein*	.381	.156	11	.673	.397	.185

Signifikanz: ** $p < .05$; * $p < .10$

(Quelle: eigene Darstellung)

Wird für das Konfliktverhältnis mit der Variable *Konflikthäufigkeit* gerechnet, zeigt der Korrelationskoeffizient von Kendall-Tau-b (r_{tb}) mit dem Wert von .372 ($p = .159$, $n = 12$) bereits ein etwas anderes Bild. Der positive Korrelationskoeffizient verzeichnet eine höhere Effektstärke, jedoch ohne statistische Signifikanz. Der exakte Test nach Fisher und die symmetrischen Tau-c Koeffizienten bestätigen auch hier, dass zwischen der *Konflikthäufigkeit* und der *destruktiven Beeinflussungsart* ($p = .945$; $\tau_c = .313$, $p = .183$) kein statistisch signifikanter Zusammenhang vorliegt. Die Effektstärke des Kendall-Tau-b liegt bei .381 ($p = .156$, $n = 11$) und ist nicht statistisch signifikant. Gemäss dem exakten Test nach Fisher und den Tau-c-Koeffizeiten existiert zwischen der *Konflikthäufigkeit* und der *destruktiven Beeinflussungsart Allgemein* ($p = .673$; $\tau_c = .397$, $p = .185$) kein statistisch signifikanter Zusammenhang. Auch der positive Korrelationskoeffizient ist trotz der höheren Effektstärke nicht signifikant ($r_{tb} = .381$, $p = .156$, $n = 11$). Insgesamt verzeichnet die *Konflikthäufigkeit* bessere Werte in Bezug auf die Effektstärken und Signifikanzniveaus der Rangkorrelation von Kendall-Tau-b und des Chi-Quadrat-Tests mittels Kendall-Tau-c. Die berechneten Zusammenhangsmasse liefern Hinweise darauf, dass ein konfliktgeprägtes Verhältnis in positivem Zusammenhang mit der destruktiven Beeinflussungsart steht. Aufgrund der fehlenden statistischen Signifikanz wird H1 abgelehnt.

Nachfolgend wird der in H4 postulierte Zusammenhang zwischen dem konfliktgeprägten Verhältnis und dem *Anreizsystem* sowie dem *direkten Einfluss* untersucht. Die Variable *Anreizsystem* ist rechtsschief verteilt (Schiefe: .840) und zeigt, dass 56 Prozent der Auftraggeber ($n = 10$) generell noch nie mit einem Anreizsystem auf die Unzufriedenheit mit einer Evaluation reagiert haben. 39 Prozent ($n = 7$) haben dies schon einmal und nur eine

Person (6%) mehr als einmal gemacht (*N* = 18). Bei der stark rechtsschiefen Variable (Schiefe: 1.613) des *direkten Einflusses* zeigt sich noch ein extremeres Bild, wobei sogar 72 Prozent der Auftraggeber (*n* = 13) noch nie einen direkten Einfluss ausgeübt haben. Nur drei Personen (17%) haben dies einmal und zwei Personen (11%) mehr als einmal gemacht (*N* = 18). Insgesamt hat kein Auftraggeber den Evaluierenden jemals in Aussicht gestellt, das Honorar nicht zu zahlen (*N* = 18). Die Massnahme innerhalb des Anreizsystems mit den meisten Nennungen liegt darin, dass den Evaluierenden mitgeteilt wird sie nicht mehr für zukünftige Aufträge zu berücksichtigen. 22 Prozent der Auftraggeber (*n* = 4) haben dies mindestens einmal gemacht. Jeweils 17 Prozent (*n* = 3) haben den Evaluierenden erklärt, dass der Auftrag entzogen werden könnte oder es wurden klare Anreize für Ergebnissänderungen nahegelegt. Den Evaluierenden zu drohen den Bericht nicht zu veröffentlichen und sie aufzufordern Ergebnisse abzuändern währendem ihnen die negativen Folgend für ihr Vorgehen verdeutlicht werden, wurde jeweils von 11 Prozent der Auftraggeber (*n* = 2) mindestens einmal gemacht (*N* = 18). Zur Überprüfung der zweigeteilten H4, ob ein positiver Zusammenhang zwischen dem konfliktgeprägten Verhältnis und dem *Anreizsystem* (H4a) resp. dem *direkten Einfluss* (H4b) besteht, wurden Rangkorrelationen und Chi-Quadrat-Tests gerechnet (siehe Tabelle 4).

Tabelle 4: Übersicht der Resultate zur Überprüfung von H4

	Rangkorrelation			Chi-Quadrat-Tests		
	Kendall-Tau-b			Exakter Test nach Fisher	Kendall-Tau-c	
	Korrelationskoeffizient r_b	Signifikanz (2-seitig)	*N*	Exakte Signifikanz (2-seitig)	Wert τ_c	Exakte Signifikanz
Konfliktverhältnis x *Anreizsystem*	.215	.308	18	.030**	.213	.325
Konflikthäufigkeit x *Anreizsystem*	.459	.092*	11	.706	.471	.105
Konfliktverhältnis x *direkter Einfluss*	.258	.216	18	.363	.231	.234
Konflikthäufigkeit x *direkter Einfluss*	.404	.135	11	.855	.397	.151

Signifikanz: ** $p < .05$; * $p < .10$

(Quelle: eigene Darstellung)

Die Korrelationskoeffizienten von Kendall-Tau-b fallen bei der rechnerischen Überprüfung von H4a und H4b mit der Variable *Konfliktverhältnis* (*N* = 18) sowie der Variable *Konflikthäufigkeit* (*N* = 11) statistisch nicht signifikant, aber gemäss der Annahme positiv aus. Auch die Ergebnisse des Chi-Quadrat-Test Kendall-Tau-c postulieren einen positiven Zusammenhang für die Variablen beider Teilhypothesen. Gemäss dem exakten Test nach Fisher besteht ein positiver, statistisch signifikanter Zusammenhang zwischen dem *Konfliktverhältnis* und dem *Anreizsystem* (*p = .030***). Der Tau-c-Koeffizient gibt jedoch kein statistisch signifikanter Zusammenhang zwischen dem *Konfliktverhältnis* und dem

Anreizsystem (τ_c = .213, p = .325) an. Wird stattdessen mit der *Konflikthäufigkeit* als unabhängige Variable gerechnet, ergibt weder der exakte Test nach Fisher (p = .706) noch der Chi-Quadrat-Test mittels Kendal-Tau-c (τ_c = .471, p = .105) ein statistisch signifikantes Resultat. Der Korrelationskoeffizient von Kendall-Tau-b (r_{tb}) fällt jedoch mit dem Wert von .459 (p = .092*, n = 11) und damit einer höheren Effektstärke statistisch signifikant aus. Die gefundene positive, statistisch signifikante Zusammenhänge zwischen dem *Konfliktverhältnis* sowie der *Konflikthäufigkeit* und dem *Anreizsystem* führen dazu, dass der erste Teil der H4 somit vorläufig bestätigt wird. Wird der Zusammenhang zwischen einem konfliktgeprägten Verhältnis und dem *direkten Einfluss* berechnet, ergeben sowohl der exakte Test nach Fisher (*Konfliktverhältnis: p* = .363; *Konflikthäufigkeit: p* = .855) als auch Kendall-Tau-c (*Konfliktverhältnis: p* = .234; *Konflikthäufigkeit: p* = .151) keine statistisch signifikanten Werte. Auch die Rangkorrelationskoeffizienten nach Kendall-Tau-b fallen statistisch nicht signifikant aus (*Konfliktverhältnis:* r_{tb} = .258, p = .216, n = 18; *Konflikthäufigkeit:* r_{tb} = .404, p = .135, n = 11). Insgesamt ergeben die Berechnungen der Zusammenhangsmasse mit der Variable *Konflikthäufigkeit* vergleichsweise grössere Effektstärken und tiefere Signifikanzniveaus als mit der Variable *Konfliktverhältnis* (vergleiche auch Tabelle 3). Obwohl die Ergebnisse auf einen positiven Zusammenhang hinweisen, konnte nicht statistisch signifikant bestätigt werden, dass ein konfliktgeprägtes Verhältnis tatsächlich mit einem stärkeren direkten, expliziten Einfluss der Auftraggeber auf Evaluierende zusammenhängt. Folglich wird der zweite Teil von H4 abgelehnt.

4.1.3 Unzufriedenheit

Insgesamt waren 57 Prozent (n = 20; N = 35) jemals mit einer in Auftrag gegebenen Evaluation unzufrieden. Als Gründe der Unzufriedenheit haben 81 Prozent der Auftraggeber (n = 17; N = 21) die ungenügende Qualität der Evaluation und 57 Prozent (n = 12) die Nichterfüllung von Erwartungen angegeben. Für jeweils 29 Prozent (n = 6) lagen die Gründe der Unzufriedenheit darin, dass der Zeitplan nicht eingehalten wurde und die Evaluierenden den Kontext der Evaluation nicht verstanden haben. Die zusätzliche offene Frage offenbarte weitere Gründen der Unzufriedenheit. Ein Respondent gab an, dass die „Komplexität der Programmstruktur und Interessen der Förderer die Evaluation erschwerten“, weiter wurde erwähnt, dass „der Evaluationsbericht die Stärken und Erfolge des Projekts nicht so sehr betont“ hätte, wie es sich der Auftraggeber gewünscht hätte. Als weitere Ursachen für die Unzufriedenheit wurde die mangelnde Qualität der Analyse

und Synthese, der Erfahrungsmangel eines Evaluierenden im Evaluationsfeld und die Unfähigkeit die Evaluation mit der Strategie zu verknüpfen, genannt. Zudem wurde aufgeführt, dass „der Evaluierende die Ergebnisse des Berichts, die während der mündlichen Besprechung erläutert wurden, gedämpft hatte“. Die minim linksschiefe Variable *Unzufriedenheit* (Schiefe: - .045), welche die Intensität der Unzufriedenheit misst, besitzt zwei Modi, wobei 32 Prozent der Auftraggeber (n = 6; N = 19) insgesamt *ziemlich zufrieden* mit unterschiedlichen Aspekten derjenigen Evaluation sind, in Bezug auf welche sie während der Umfrage ihre Unzufriedenheit äusserten. Der zweite Modus zeigt, dass ebenfalls 32 Prozent in der gesamtheitlichen Einschätzung einer solchen Evaluation *ziemlich unzufrieden* (n = 6) sind. Für die Aspekte der methodischen Vorgehensweise und der Evaluations- und Methodenkompetenz der Evaluierenden ordneten jeweils 21 Prozent der Auftraggeber (n = 4; N = 19) ihre Unzufriedenheit dem Wert 7 resp. 8 (von insgesamt 10) zu. Die Unzufriedenheit hinsichtlich der Einhaltung des Zeitplans bewerteten 22 Prozent (n = 4; N = 18) und 17 Prozent (n = 3; N = 18) hinsichtlich der Evaluationsqualität jeweils mit dem Wert 8 auf der Unzufriedenheitsskala.

Die in Tabelle 5 dargestellten Ergebnisse geben Hinweise darauf, ob ein positiver Zusammenhang zwischen der *Unzufriedenheit* und der *destruktiven Beeinflussungsart* der Auftraggeber vorliegt (H2).

Tabelle 5: Übersicht der Resultate zur Überprüfung von H2

	Rangkorrelation			Chi-Quadrat-Tests		
	Kendall-Tau-b			Exakter Test nach Fisher	Kendall-Tau-c	
	Korrelationskoeffizient r_b	Signifikanz (2-seitig)	N	Exakte Signifikanz (2-seitig)	Wert τ_c	Exakte Signifikanz
Unzufriedenheit x *destruktive Beeinflussungsart*	.205	.433	13	.021**	.213	.573
Unzufriedenheit x *destruktive Beeinflussungsart Allgemein*	.190	.421	14	.347	.177	.444

Signifikanz: ** $p < .05$; * $p < .10$

(Quelle: eigene Darstellung)

Gemäss der Rangkorrelationskoeffizienten von Kendall-Tau-b ist weder der Zusammenhang der *Unzufriedenheit* mit der *destruktiven Beeinflussungsart* (r_{tb} = .205, p = .433, n = 13), noch mit ersterer und der *destruktiven Beeinflussungsart Allgemein* (r_{tb} = .190, p = .421, n = 14) statistisch signifikant. Der exakte Test nach Fisher zeigt jedoch für den positiven Zusammenhang der *Unzufriedenheit* mit der *destruktiven Beeinflussungsart* mit einem p-Wert von .021** ein statistisch signifikantes Ergebnis. Der Zusammenhang mit der *destruktiven Beeinflussungsart Allgemein* als abhängige Variable ist statistisch nicht signifikant (p = .347). Die auf dem Chi-Quadrat-Test basierenden Werte von Kendall-

Tau-c sind bei beiden Zusammenhängen nicht signifikant, wobei Kendall-Tau-c bei der abhängigen Variable der *destruktiven Beeinflussungsart* bei .213 (p = .573) und bei der *destruktiven Beeinflussungsart Allgemein* bei .177 (p = .444) liegt. Die Werte der Rangkorrelationskoeffizienten und von Kendall-Tau-c weisen alle auf einen positiven Zusammenhang hin, der durch das signifikante Testergebnis zusätzlich unterstrichen wird. Folglich wird H2 vorläufig bestätigt.

4.1.4 Schwierigkeiten

Die rechtsschiefe Variable *Schwierigkeiten Häufigkeit* weist bezüglich ihrer Schiefe (.623) auf eine gesamtheitlich tiefe Häufigkeit von Schwierigkeiten in der Zusammenarbeit mit Evaluierenden hin. Dies zeigt sich darin, dass sich 67 Prozent (n = 20; N = 30) der Aufraggeber von nie bis eher selten mit Schwierigkeiten konfrontiert sehen. Dabei lässt sich die Häufigkeit von Schwierigkeiten für 10 Prozent (n = 3) als nie, für 30 Prozent (n = 9) der Auftraggeber als sehr selten und für 27 Prozent (n = 8) als eher selten einordnen. Als Schwierigkeiten wurden insb. angegeben, dass den Evaluierenden das Verständnis für die zu evaluierende Organisation (51%, n = 18; N = 35) sowie das gegenseitige Verständnis zwischen dem Evaluierenden und den Auftraggebern (46%, n = 16; N = 35) fehlt. Das Fehlen von wichtigen Fachkompetenzen resp. Ressourcen gaben 44 Prozent (n = 15; N = 34) resp. 41 Prozent (n = 13; N = 32) der Auftraggeber an. Dass die Unabhängigkeit von Evaluationen schwierig aufrechtzuerhalten ist, ohne die Evaluierenden zu stark zu beeinflussen wurde von 15 Prozent (n = 9; N = 34) bejaht und 9 Prozent (n = 3; N = 33) zählten unmotivierte Evaluierende zu den aufgetretenen Schwierigkeiten während ihrer Zusammenarbeit. Bei der offenen Frage wurden u.a. auf die Herausforderung hingewiesen, „Evaluierende zu finden, die sowohl über kontextuelle Expertise in die Perspektive der Programmbegünstigten als auch über methodische Expertise verfügen“. Neben Schwierigkeiten aufseiten der Auftraggeber wie Sprachbarrieren, mangelnder Zeit die Evaluierenden gut zu managen oder „die Evaluierenden ausreichend zu integrieren, ohne die Organisation zu belasten“, wurden ebenso auf Schwierigkeiten seitens der Evaluierenden hingewiesen. Eine Person gab an, dass es schwierig sei das Programmwachstum in der Evaluation ausreichend zu berücksichtigen, „da sich Evaluierende auf das aktuelle Paradigma beschränken und nicht verstehen wie in einem Umfeld von schnellem Wachstum und Innovation gedacht und evaluiert werden soll“. Zudem wurde beschrieben, dass „einige Evaluierende nur theoretische Modellnutzer sind, aber tatsächliche Geschäftsvorgänge nicht kennen“ oder nicht im Prozess der Datenerhebung involviert sein wollen.

Im Rahmen von H3 wird der Zusammenhang zwischen der Häufigkeit von *Schwierigkeiten,* die Auftraggeber in der Zusammenarbeit mit den Evaluierenden wahrnehmen und der *destruktiven Beeinflussungsart* untersucht (siehe Tabelle 6).

Tabelle 6: Übersicht der Resultate zur Überprüfung von H3

	Rangkorrelation			Chi-Quadrat-Tests		
	Kendall-Tau-b			Exakter Test nach Fisher	Kendall-Tau-c	
	Korrelationskoeffizient r_b	Signifikanz (2-seitig)	*N*	Exakte Signifikanz (2-seitig)	Wert τ_c	Exakte Signifikanz
Schwierigkeiten x *destruktive Beeinflussungsart*	.162	.381	23	.878	.153	.394
Schwierigkeiten x *destruktive Beeinflussungsart Allgemein*	-.151	.385	25	.438	-.141	.398

Signifikanz: ** $p < .05$; * $p < .10$

(Quelle: eigene Darstellung)

Für den Zusammenhang zwischen den *Schwierigkeiten* und der *destruktiven Beeinflussungsart* ergibt sich kein statistisch signifikantes Ergebnis. Der Korrelationskoeffizient von Kendall-Tau-b liegt bei .162 ($p = .381$, $n = 23$) und weist analog zum symmetrischen Mass des Kendall-Tau-c (τ_c) mit dem Wert von .153 ($p = .394$) auf einen positiven, aber nicht statistisch signifikanten Zusammenhang hin. Beim exakten Test von Fisher ergibt sich ein nicht statistisch signifikanter p-Wert von .878. Für den statistisch nicht signifikanten Zusammenhang der *Schwierigkeiten* und der *destruktiven Beeinflussungsart Allgemein* zeigt der Zusammenhang beim Rangkorrelationskoeffizienten ($r_{tb} = -.151$, $p = .385$, $n = 25$) sowie dem Kendall-Tau-c ($\tau_c = -.141$, $p = .398$) in die gegensätzliche resp. negative Richtung. Gemäss dem exakten Test nach Fisher liegt ebenfalls kein statistisch signifikantes Ergebnis vor ($p = .438$). Neben der fehlenden statistischen Signifikanz der Zusammenhänge weisen die Ergebnisse in Bezug auf die mögliche Richtung eines Zusammenhangs eine gewisse Ambivalenz auf. Folglich wird H3 abgelehnt.

4.2 Resultate zu den Eigenschaften der Auftraggeber

Der Zusammenhang der Eigenschaften der Auftraggeber – wie die *Vertrautheit* mit Standards und die *Berufserfahrung* – als unabhängige Variablen wird mittels der sechsten und siebten Hypothesen mit der *konstruktiven Beeinflussungsart* als abhängige Variable überprüft. Bei der fünften Hypothese wird geprüft, ob die *Vertrautheit* erneut als unabhängige Variable mit der abhängigen Variable der *Erwartung an Unabhängigkeit* zusammenhängt.

4.2.1 Konstruktive Beeinflussungsart

Bei der linksschief-verteilten (Schiefe: -.477) Variable *konstruktive Beeinflussungsart* haben alle Auftraggeber ($N = 23$) Änderungsvorschläge geäussert, welche der konstruktiven Beeinflussungsart entsprechen (siehe Tabelle 7). Neun Respondenten (39%) haben dabei schon mehr als einmal eine konstruktive Beeinflussungsart ausgeübt, wobei dies 14 Auftraggeber (61%) als Mehrheit schon oft gemacht haben.

Tabelle 7: Übersicht der Resultate zur konstruktiven Beeinflussungsart

	0	1	2	3	*n*
Diskussion der Evaluationsmethoden zur Verbesserung der Evaluationsqualität	3.7 (1)	14.8 (4)	51.9 (14)	29.6 (8)	27
Proaktive Identifikation von Verbesserungspotenzial im Evaluationsprozess	11.1 (3)	7.4 (2)	37.0 (10)	44.4 (12)	27
Illustration von Verbesserungspotenzial ohne Verzerrung der Evaluationsresultate	14.8 (4)	7.4 (2)	37.0 (10)	40.7 (11)	27
Diskussion der Ergebnispräsentation zur Verbesserung des Zielgruppenverständnis	3.8 (1)	11.5 (3)	42.3 (11)	42.3 (11)	26
Ideenaustausch zur Qualitätsverbesserung	3.7 (1)	7.4 (2)	40.7 (11)	48.1 (13)	27
Neutrale Diskussion der Schlussfolgerungen	3.8 (1)	-	38.5 (10)	57.7 (15)	26
Konstruktiver Dialog über Verbesserung der Evaluationsqualität	-	4.0 (1)	32.0 (8)	64.0 (16)	25
Unterstützung durch frühzeitige Informationslieferung	-	-	19.2 (5)	80.8 (21)	26
Konstruktive Beeinflussungsart	-	-	39.1 (9)	60.9 (14)	23
Konstruktive Beeinflussungsart Allgemein	-	10.3 (3)	31.0 (9)	58.6 (17)	29
Betterment	6.9 (2)	17.2 (5)	31.0 (9)	44.8 (13)	29
Support	-	9.7 (3)	25.8 (8)	64.5 (20)	31

Anmerkungen: Skala von 0 (= Nein, das habe ich nie gemacht), 1 (= Ja, das habe ich einmal gemacht), 2 (= Ja, das habe ich mehr als einmal gemacht), 3 (= Ja, das habe ich oft gemacht); *n* = Gesamtzahl aller valider Fälle. Die Zahlen repräsentieren Prozentwerte, in Klammern wurden die jeweiligen Häufigkeiten angegeben.

(Quelle: eigene Darstellung)

Alle Auftraggeber ($N = 26$) haben angegeben, dass sie die Arbeit der Evaluierenden unterstützen, indem sie frühzeitig alle für die Evaluation relevanten Informationen geliefert haben. 81 Prozent ($n = 21$) der Aufraggeber haben angegeben, dies oft gemacht zu haben, wobei 19 Prozent ($n = 5$) dies schon mehr als einmal gemacht haben. Ebenso haben alle Auftraggeber ($N = 25$) angegeben, dass sie mit den Evaluierenden schon mindestens einmal ein konstruktiver Dialog darüber geführt haben wie die Evaluationsqualität verbessert werden kann. Eine Person (4%) hat dies einmal, acht Personen (32%) mehr als einmal und 16 Personen (64%) oft gemacht. Der Aussage, dass mit Evaluierenden Schlussfolgerungen neutral diskutiert (96%, $n = 25$; $N = 26$), Ideen zur Verbesserung der Evaluationsqualität ausgetauscht wurden (96%, $n = 26$; $N = 27$) sowie diskutiert wurde, welche Evaluationsmethoden zur Verbesserung der Evaluationsqualität eingesetzt (96%, $n = 26$; $N = 27$) und wie die Resultate zur Verbesserung des Zielgruppenverständnis präsentiert werden können (96%, $n = 25$; $N = 26$), haben alle Respondenten mit Ausnahme jeweils einer

Person zugestimmt. Die proaktive Identifikation von Verbesserungspotenzial im Evaluationsprozess haben drei Befragte (11%) noch nie gemacht, wobei dies 2 Personen (7%) einmal, zehn Personen (37%) mehr als einmal und zwölf Personen (44%) oft gemacht haben (N = 27). Auch die proaktive Illustration von Verbesserungspotenzial ohne Verzerrung der Evaluationsresultate haben vier Auftraggeber (15%) noch nie gemacht. Demgegenüber wurde dies von zwei Personen (7%) bereits einmal, von zehn Personen (37%) mehr als einmal und von elf Personen (41%) oft gemacht (N = 27). Gemäss der linksschiefen Variable *konstruktive Beeinflussungsart Allgemein* (Schiefe: -.996) haben alle Auftraggeber (N = 29) generell einen konstruktiven Einfluss auf Evaluierende ausgeübt. Darunter haben im Allgemeinen jeweils drei Personen (10%) einmal, neun Personen (31%) mehr als einmal und 17 Personen (59%) schon oft den Evaluationsprozess auf eine konstruktive Art beeinflusst. Für die Variable *Betterment* als direkter, konstruktiver Beeinflussungstyp zeigt sich, dass 93 Prozent (n = 27) der Auftraggeber mindestens einmal den Evaluierenden gezeigt haben, welche Punkte verbessert werden können, um eine bessere Evaluationsqualität zu erreichen, ohne dabei die Resultate zu verzerren. Fünf Personen (17%) haben dies einmal, neun Personen (31%) mehr als einmal und 13 Personen (45%) oft gemacht. Demgegenüber haben dies lediglich zwei Auftraggeber (7%) noch nie gemacht (N = 29). Beim indirekten, konstruktiven Beeinflussungstyp *Support* zeigt sich ein ähnliches Muster, jedoch mit Gewicht auf die stärkere Ausübungshäufigkeit dieser Beeinflussungsart. Alle Auftraggeber (N = 31) geben nämlich an, sich generell für die Optimierung der Evaluationsqualität einzusetzen. Die Häufigkeiten variieren insofern, dass dies jeweils drei Personen (10%) einmal, acht Personen (26%) mehr als einmal und 20 Personen als Mehrheit (65%) oft gemacht haben. Als Spiegelbild der destruktiven Beeinflussungsart zeigt die Variable *Beeinflussungsintention* für keinen Auftraggeber eine konstruktive Beeinflussungsform an (N = 26) und nimmt lediglich Werte bei den Ausprägungen hinsichtlich der destruktiven Beeinflussungsart ein (siehe Kapitel 4.1.1).

4.2.2 Vertrautheit

Insgesamt 70 Prozent (n = 21; N = 30) aller Auftraggeber kennen die nationalen „Program Evaluation Standards“ nicht. Lediglich 30 Prozent (n = 9) aller Respondenten geben ihre *Standardkenntnis* an. Bei der minim linksschief-verteilten Variable *Vertrautheit* (Schiefe: -.018) sind insgesamt sechs Auftraggeber mit den nationalen Evaluationsstandards eher gut vertraut (67%), wobei damit zwei Personen (22%) eher schlecht vertraut sind und eine Person (11%) sehr gut vertraut ist (N = 9). In Bezug auf die Wichtigkeit dieser Evaluationsstandards ordnen sieben Personen (78%) die *Standardwichtigkeit* als eher wichtig

ein, wobei jeweils eine Person (11%) die Standards als eher unwichtig und sehr wichtig (11%) ansieht.

Im Rahmen der H5 wird der Zusammenhang zwischen der *Vertrautheit* und der *Erwartung an die Unabhängigkeit* von Evaluationen untersucht (siehe Tabelle 8). Für weitere Angaben zur *Erwartung an die Unabhängigkeit* wird auf Tabelle 12 verwiesen.

Tabelle 8: Übersicht der Resultate zur Überprüfung von H5

	Rangkorrelation			Chi-Quadrat-Tests		
	Kendall-Tau-b			Exakter Test nach Fisher	Kendall-Tau-c	
	Korrelationskoeffizient r_b	Signifikanz (2-seitig)	*N*	Exakte Signifikanz (2-seitig)	Wert τ_c	Exakte Signifikanz
Vertrautheit x *Erwartung Unabhängigkeit*	.169	.666	7	.619	.163	.905

Signifikanz: ** $p < .05$; * $p < .10$

(Quelle: eigene Darstellung)

Für den Zusammenhang zwischen der *Vertrautheit* und der *Erwartung an die Unabhängigkeit* zeigt sich kein statistisch signifikantes Ergebnis. Der Korrelationskoeffizient von Kendall-Tau-b deutet einen positiven Zusammenhang an und liegt bei .169 (p = .666, n = 7). Beim exakten Test von Fisher liegt der p-Wert bei .619 und der Wert von Kendall-Tau-c bei .163 (p = 905), wobei letzterer ebenso auf einen positiven Zusammenhang der Variablen hinweist. Obwohl die Ergebnisse auf einen positiven Zusammenhang hinweisen, konnte nicht statistisch signifikant bestätigt werden, dass die Vertrautheit der Auftraggeber mit den Evaluationsstandards tatsächlich mit einer erhöhten Erwartung an die Unabhängigkeit von Evaluationen zusammenhängt. Folglich wird H5 abgelehnt.

Die H6 überprüft, ob ein positiver Zusammenhang zwischen der *Vertrautheit* und der *konstruktiven Beeinflussungsart* besteht (siehe Tabelle 9).

Tabelle 9: Übersicht der Resultate zur Überprüfung von H6

	Rangkorrelation			Chi-Quadrat-Tests		
	Kendall-Tau-b			Exakter Test nach Fisher	Kendall-Tau-c	
	Korrelationskoeffizient r_b	Signifikanz (2-seitig)	*N*	Exakte Signifikanz (2-seitig)	Wert τ_c	Exakte Signifikanz
Vertrautheit x *konstruktive Beeinflussungsart*	.522	.186	7	1.000	.490	.571
Vertrautheit x *konstruktive Beeinflussungsart Allgemein*	.639	.052*	9	.429	.519	.087*

Signifikanz: ** $p < .05$; * $p < .10$

(Quelle: eigene Darstellung)

Die Korrelationskoeffizienten von Kendall-Tau-b fallen bei der rechnerischen Überprüfung der *Vertrautheit* mit der Variable *konstruktive Beeinflussungsart* (r_{tb} = .522, p = .186, n = 7) sowie der Variable *konstruktive Beeinflussungsart Allgemein* (r_{tb} = .639, p = .052*,

n = 9) gemäss der Annahme positiv aus. Der positive Zusammenhang der *Vertrautheit* und der *konstruktiven Beeinflussungsart Allgemein* zeigt sich sowohl bei der Rangkorrelation von Kendall-Tau-b (*p* = .052*) als auch dem Chi-Quadrat-Test von Kendall-Tau-c (τ_c = .519, *p* = .087*) als statistisch signifikant. Gemäss Cohen (1988) deuten die Werte des Rangkorrelationskoeffizienten von .639 und des Kendall-Tau-c von .519 auf starke Effektgrössen hin. Lediglich der exakte Test nach Fisher ist mit einem p-Wert von .429 nicht signifikant. Beim Zusammenhang zwischen der *Vertrautheit* und der *konstruktiven Beeinflussungsart* liefern die Chi-Quadrat-Tests mittels des exakten Test nach Fisher einen p-Wert von 1.000 und mittels Kendall-Tau-c einen Wert von .490 (*p* = .571), die beide nicht statistisch signifikant sind. Der anhand zweier Tests gefundene positive, statistisch signifikante Zusammenhang zwischen der *Vertrautheit* und der *konstruktiven Beeinflussungsart Allgemein* führt dazu, dass H6 somit vorläufig bestätigt wird.

4.2.3 Berufserfahrung

Werden die rechtsschiefen Verteilungen der Variablen *Berufserfahrung, Evaluationsjahre* und *Evaluationsanzahl* miteinander verglichen, fällt auf, dass die Variable *Evaluationsjahre* die grösste (Schiefe: .888) und die Variable *Evaluationsanzahl* die kleinste Schiefe (.272) aufweist. Die *Berufserfahrung* wurde als additiver Index aus den beiden Variablen gebildet und lässt sich somit in der Mitte einordnen. Bei der Variable *Evaluationsjahre* haben bereits 57 Prozent der Auftraggeber (*n* = 32*)* bis zu fünf Jahre Erfahrung in der Vergabe von Evaluationsaufträgen. 27 Prozent der Auftraggeber (*n* = 15) weisen sogar sechs bis 10 Jahre an Erfahrung auf, wobei lediglich 5 Prozent (*n* = 3) 11 bis 15 Jahre und 11 Prozent (*n* = 6) seit mehr als 15 Jahren Erfahrungen in diesem Bereich gesammelt haben (*N* = 56). Bei der Variable *Evaluationsanzahl* haben insgesamt 76 Prozent (*n* = 41) der Auftraggeber bis zu 20 Evaluationen in ihrer bisherigen Karriere durchgeführt bzw. waren daran als Auftraggeber beteiligt. Davon haben 39 Prozent (*n* = 21) eine bis fünf Evaluationen und 37 Prozent (*n* = 20) sechs bis 20 Evaluationen durchgeführt. Zudem haben 24 Prozent der Auftraggeber (*n* = 13) schon mehr als 20 Evaluationen durchgeführt (*N* = 54). Bei der Variable *Berufserfahrung* zeigt sich, dass über 56 Prozent der Auftraggeber (*n* = 30) eher unerfahren sind und sich gegenüber dieser Mehrheit 26 Prozent der Auftraggeber (*n* = 14) von eher erfahren (15%, *n* = 8)*,* über sehr erfahren *(*6%, *n* = 3) bis äusserst erfahren *(*6%, *n* = 3) einordnen lassen (*N* = 54). Die H7 überprüft, ob ein negativer Zusammenhang zwischen der *Berufserfahrung* und der *konstruktiven Beeinflussungsart* besteht. Die Korrelationskoeffizienten von Kendall-Tau-b fallen bei

der rechnerischen Überprüfung der *Berufserfahrung* mit der Variable *konstruktive Beeinflussungsart* (r_{tb} = .323, p = .092*, n = 23) sowie der Variable *konstruktive Beeinflussungsart Allgemein* (r_{tb} = .253, p = .119, n = 29) entgegen der Annahme positiv aus (siehe Tabelle 10).

Tabelle 10: Übersicht der Resultate zur Überprüfung von H7

	Rangkorrelation			Chi-Quadrat-Tests		
	Kendall-Tau-b			Exakter Test nach Fisher	Kendall-Tau-c	
	Korrelationskoeffizient r_b	Signifikanz (2-seitig)	*N*	Exakte Signifikanz (2-seitig)	Wert τ_c	Exakte Signifikanz
Berufserfahrung x *konstruktive Beeinflussungsart*	.323	.092*	23	.746	.393	.101
Berufserfahrung x *konstruktive Beeinflussungsart Allgemein*	.253	.119	29	.647	.253	.123
Evaluationsjahre x *konstruktive Beeinflussungsart*	.352	.078*	23	.193	.383	.084*
Evaluationsjahre x *konstruktive Beeinflussungsart Allgemein*	.268	.114	29	.102	.246	.121
Evaluationsanzahl x *konstruktive Beeinflussungsart*	.272	.178	23	.172	.302	.189
Evaluationsanzahl x *konstruktive Beeinflussungsart Allgemein*	.292	.090*	29	.508	.264	.094*

Signifikanz: ** $p < .05$; * $p < .10$

(Quelle: eigene Darstellung)

Der positive Zusammenhang der *Berufserfahrung* und der *konstruktiven Beeinflussungsart* zeigt sich sowohl bei der Rangkorrelation von Kendall-Tau-b (r_{tb} = .323, p = .092*, n = 23) als statistisch signifikant und beim Chi-Quadrat-Test von Kendall-Tau-c (τ_c = .393, p = .101) als knapp nicht statistisch signifikant. Gemäss Cohen (1988) deuten die Werte des Rangkorrelationskoeffizienten von .323 und des Kendall-Tau-c von .393 auf mittlere Effektgrössen hin. Der exakte Test nach Fisher ist mit einem p-Wert von .746 jedoch nicht signifikant. Beim Zusammenhang zwischen der *Berufserfahrung* und der *konstruktiven Beeinflussunsgart Allgemein* liefern die Chi-Quadrat-Tests mittels des exakten Test nach Fisher einen nicht signifikanten p-Wert von .647 und mittels Kendall-Tau-c einen Wert von .253 (p = .123). Werden die Zusammenhänge der *Evaluationsjahre* und der *Evaluationsanzahl* mit der abhängigen Variable der *konstruktiven Beeinflussungsart (Allgemein)* untersucht, zeichnet sich ein ähnliches Muster ab. Für die *Evaluationsjahre* und die *konstruktive Beeinflussungsart* zeigen die Rangkorrelation nach Kendall-Tau-b (r_{tb} = .352, p = .078*, n = 23) und der Chi-Quadrat-Test nach Kendall-tau-c (τ_c = .383, p = .084*) ein positiver, statistisch signifikanter Zusammenhang. Nach Cohen (1988) sind dabei beide Effektgrössen mittelstark ausgeprägt. Auch für die *Evaluationsanzahl* und die *konstruktive Beeinflussungsart Allgemein* liegt ein positiver, statistisch signifikanter

Zusammenhang vor, was durch den Rangkorrelationskoeffizient mit einem Wert von .292 (p = .090*, n = 29) und dem Wert von Kendall-Tau-c mit .264 (p = .094*) gezeigt wird. Die Effektgrössen sind dabei etwas schwächer ausgeprägt, liegen aber immer noch im mittelstarken Bereich. Die p-Werte des exakten Tests nach Fisher fallen bei allen Zusammenhängen nicht statistisch signifikant aus. Beim Zusammenhang zwischen den *Evaluationsjahren* und der *konstruktiven Beeinflussungsart Allgemein* liegt beinahe ein statisch signifikanten p-Wert vor, der einer Irrtumswahrscheinlichkeit von α mit 10% entspricht. Die anhand dieser Tests gefundenen positiven, statistisch signifikanten Zusammenhänge zwischen der *Berufserfahrung* – inklusive *Evaluationsjahre* und *Evaluationsanzahl* – und der *konstruktiven Beeinflussungsart* weisen auf die der Hypothese entgegengesetzte Richtung hin. Folglich wird die H7 trotz signifikanter Resultate abgelehnt.

4.3 Ergänzende Resultate

Zur Beantwortung der Forschungsfragen wird mittels der nachfolgenden Resultate ergänzend auf die Wichtigkeit und Wahrnehmung von unabhängigen Evaluationen sowie auf die Konfliktgründe und die vorgeschlagenen präventiven Massnahmen seitens der US-Auftraggeber eingegangen.

4.3.1 Wichtigkeit und Wahrnehmung der Unabhängigkeit von Evaluationen

In diesem Abschnitt wird einerseits die Wichtigkeit der Unabhängigkeit von Evaluationen beschrieben. Andererseits werden die Resultate zu den Wahrnehmungen in Bezug auf die generelle Unabhängigkeit von Evaluationen und der eigenen Einflussstärke erläutert. Für die Mehrheit der Auftraggeber (55%, n = 17; N = 31) wird die Unabhängigkeit von Evaluationen, d.h. dass Evaluationen generell ohne Einfluss aufseiten der Auftraggeber durchgeführt werden, als sehr wichtig eingestuft (*Wichtigkeit Unabhängigkeit)*. Insgesamt ordnen 13 Prozent der Auftraggeber (n = 4) die Unabhängigkeit von Evaluationen als tendenziell unwichtig (Werte von 0 bis 2) und 87 Prozent (n = 27) als tendenziell wichtig (Werte von 3 bis 5) ein. Die linksschiefe Variable *Wahrnehmung der Unabhängigkeit*[10] (Schiefe: -0.391) zeigt, dass alle Auftraggeber (N = 30) die generell wahrgenommene Unabhängigkeit von durchgeführten Evaluationen den Werten 3 bis 5 auf der Skala zuordnen. 43 Prozent aller Respondenten (n = 13) ordnen die wahrgenommene Unabhängigkeit der durchgeführten Evaluationen als sehr unabhängig ein, wobei 33 Prozent (n =

[10] *Wahrnehmung der Unabhängigkeit:* Skala von 0 (= überhaupt nicht unabhängig) bis 5 (sehr unabhängig)

10) die Unabhängigkeit mit dem Wert 4 resp. 23 Prozent (n = 7) mit dem Wert 3 einordnen. Die von den Auftraggebern durchschnittlich wahrgenommene *Einflussstärke*[11], die sie auf den Evaluationsprozess resp. auf die Evaluierenden ausüben, liegt bei einem Wert von 4.60 (SD = 2.7; N = 30). Obwohl über die Mehrheit der Auftraggeber den eigenen Einfluss auf den Evaluationsprozess und die Evaluierende mit den Werten von 0 bis 5 eingeordnet haben (67%, n = 20), haben trotzdem 17 Prozent der Auftraggeber (n = 5) ihren eigenen Einfluss mit dem Wert 8 eingestuft.

4.3.2 Konfliktgründe und präventive Massnahmen

Die im Kapitel 4.1.2 besprochenen Eigenschaften des *Konflikts*, des *Konfliktverhältnis* und der *Konflikthäufigkeit* werden mit Resultaten zu den Konfliktgründen ergänzt. Als Konfliktgründe wurden besonders häufig das Fehlen von wichtigen Kompetenzen aufseiten der Evaluierenden (61%, n = 11; N = 18), die mangelhafte Qualität der Evaluationsresultate (67%, n = 12) und das fehlende Verständnis der Anforderungen (72%, n = 13) genannt. Im Rahmen der offenen Frage beschrieb eine Person den „Missbrauch oder die Fehldarstellung von Befunden ausserhalb des Kontextes sowie die breite Verallgemeinerung auf Basis von kleinen und verzerrten Kohorten“ als Konfliktgrund. Zudem wurden interpersonale Kompetenzen, mangelnde kommunikative und proaktive Verhaltensweisen während dem Evaluationsprozess genannt. Bei der Variable *Unterstellung* verneinten die Auftraggeber beinahe einstimmig (n = 30; N = 32), dass ihnen noch nie von Evaluierenden unterstellt wurde, sie unter Druck gesetzt oder beeinflusst zu haben, wobei zwei Auftraggeber die „Weiss nicht“-Kategorie wählten. In Anbetracht dieser Resultate stellt sich die Frage nach geeigneten Massnahmen zur Prävention von Konflikten zwischen Auftraggeber und Evaluierenden, um ein fruchtbares Evaluationsumfeld zu schaffen (siehe Tabelle 11). Die präventive Massnahme, dass ein grösserers gegenseitiges Verständnis der Zielsetzung, Zweckbestimmung und Funktionen geschaffen werden sollte, wurde von insgesamt 82 Prozent der Auftraggeber genannt, die zumindest eine Antwort angegeben haben. Dies entspricht 23 Prozent der insgesamt gegebenen Antworten (n = 27; N = 119). Zudem haben 93 Prozent der Auftraggeber (n = 14), die jemals einen Konflikt mit einem Evaluierenden hatten diese Massnahme angegeben, wobei 77 Prozent der Auftraggeber (n = 13) ohne jemals einen Konflikt gehabt zu haben diese Antwort gewählt haben (N = 27). Unabhängig davon, ob ein *Konflikt* jemals existierte oder nicht, wurde diese präventive Massnahme somit von allen Fällen am häufigsten angegeben.

[11] *Wahrnehmung der Einflussstärke*: Skala von 0 (=kein Einfluss) bis 10 (= extrem starker Einfluss)

Tabelle 11: Übersicht der Resultate zur Prävention

Präventive Massnahmen[a]	Antworten	Fälle	Konflikt[b]		
	Prozent (N)	Prozent	Nein in Prozent (*n*)	Ja in Prozent (*n*)	*N*
Meta-Evaluation durch unabhängige Dritte	5.0 (6)	18.2	29.4 (5)	6.7 (1)	*6*
Interne oder externe Meldestelle	3.4 (4)	12.1	17.6 (3)	6.7 (1)	*4*
Engere Zusammenarbeit beider Parteien	15.1 (18)	54.5	70.6 (12)	40.0 (6)	*18*
Schaffung grösseres gegenseitiges Verständnis der Zielsetzung, Zweckbestimmung und Funktionen	22.7 (27)	81.8	76.5 (13)	93.3 (14)	*27*
Diskussion möglicher, negativer Resultate	10.9 (13)	39.4	47.1 (8)	33.3 (5)	*13*
Betonung der Verantwortung von Evaluierenden sich an Daten/Tatsachen zu halten	8.4 (10)	30.3	41.2 (7)	20.0 (3)	*10*
Verbesserung Datendokumentation	9.2 (11)	33.3	35.3 (6)	33.3 (5)	*11*
Einführung formelles Evaluationsprotokoll	12.6 (15)	45.5	52.9 (9)	40.0 (6)	*15*
Neutrale Intervention durch Dritte	5 (6)	18.2	11.8 (2)	20.0 (3)	*5*
Hinweis auf die Möglichkeit, dem publizierten Bericht eine Stellungnahme der Auftraggeber voranstellen zu lassen	5 (6)	18.2	23.5 (4)	13.3 (2)	*6*
Anderes	2.5 (3)	9.1	5.9 (1)	13.3 (2)	*3*
Gesamt	100 (119)	360.6	(17)	(15)	*32*

a. Dichotomie-Gruppe tabellarisch dargestellt bei Wert 1. Die Variable *Prävention* wurde durch die Frage gemessen: Welche der folgenden präventiven Massnahmen könnten dazu beitragen Konflikte zwischen Auftraggeber und Evaluierenden zu verhindern? Mehrfachantworten waren möglich.

b. Die dichotome Variable *Konflikt* mit den Ausprägungen Nein (= 0) und Ja (=1) misst, ob der Auftraggeber jemals mit einem Evaluator aufgrund seiner Arbeit oder Arbeitsweise einen Konflikt hatte.

(Quelle: eigene Darstellung)

Die engere Zusammenarbeit zwischen Auftraggebern und Evaluierenden wurde insgesamt von 55 Prozent der Auftraggeber als zweithäufigste Antwort genannt, die mindestens eine Antwort angegeben haben. Das sind 15 Prozent (n = 18; N = 119) der insgesamt gegebenen Antworten. Werden die Antworten in Hinblick darauf betrachtet, ob die Auftraggeber jemals einen Konflikt hatten oder nicht, fällt im Vergleich zur erstgenannten Massnahme eine umgekehrt gerichtete Verteilung auf. Dabei haben viel mehr Auftraggeber ohne Konflikt (n = 12; 71%) als mit Konflikt (n = 6; 40%) diese Massnahme angegeben. Diese Tendenz zieht sich bei den weiter genannten Massnahmen von der Einführung eines formellen Evaluationsprotokolls (*Nein:* 53%; *Ja:* 40%), der Diskussion möglicher, negativer Resultate (*Nein:* 47%; *Ja:* 33%), der verbesserten Datendokumentation (*Nein:* 35%; *Ja:* 33%), einer Meta-Evaluation durch Dritte (*Nein:* 49%; *Ja:* 7%) sowie dem Hinweis auf eine mögliche Stellungnahme (*Nein:* 24%; *Ja:* 13%) bis zur internen oder externen Meldestelle (*Nein:* 18%; *Ja:* 7%) durch. Lediglich die präventive Massnahme der neutralen Intervention durch Dritte wird häufiger von Auftraggebern mit (20%, n = 3) als ohne Konflikt (12%, n = 2) vorgeschlagen. Zum Vergleich der vorgeschlagenen präventiven Massnahmen der US-Auftraggeber und der Schweizer Auftraggeber siehe Abbildung 6.

4.4 Resultate im Ländervergleich

Die Befunde zu den Auftraggeber von Evaluationen in den USA werden mit denjenigen aus der Schweiz (Pleger & Hadorn, 2018) in Hinblick auf soziodemografische und individuelle Eigenschaften wie *Geschlecht, Berufserfahrung,* die *Evaluationsart*, der *Anteil Direktvergabe*, den *Auftraggeber-Sektor,* den *Evaluator-Sektor*, die *Erwartung an Unabhängigkeit*, *Standardkenntnis, Vertrautheit, Standardwichtigkeit*. Zudem wird das Beziehungsverhältnis von Auftraggebern und Evaluierenden der beiden Länder anhand unterschiedlicher Aspekte wie der *Reaktion,* der *Unterstellung,* der *Schwierigkeiten* und der *Prävention* verglichen.

4.4.1 Individuelle Eigenschaften der Auftraggeber im Ländervergleich

In Bezug auf die soziodemographischen und individuellen Eigenschaften der Auftraggeber lassen sich im Ländervergleich einige Gemeinsamkeiten und Unterschiede identifizieren. Während die Männer bei den Schweizer Auftraggebern überrepräsentiert sind (76%, $n = 29$, $N = 38$), sind es bei der vorliegenden Studie die weiblichen Befragten, die mit 61 Prozent ($n = 19$; $N = 31$) die Mehrheit konstituieren. Bei den Schweizer Auftraggebern betrug die durchschnittliche Anzahl Jahre, während denen Evaluationen beauftragt wurden 9.4 Jahre (SD = 7.1, N = 53). Für die USA lässt sich generell ein relativ tieferes Niveau in Bezug auf die allgemeine Berufserfahrung ausmachen, was die *Evaluationsjahre* sowie die *Evaluationsanzahl* inkludiert (siehe Kapitel 4.2.3). Hinsichtlich der Evaluationsart für welche überwiegend Aufträge vergeben werden, zeichnet sich bei beiden Ländern ein ähnliches Bild. Die US-Auftraggeber vergeben mit 83 Prozent ($n = 40$; $N = 48$) mehrheitlich externe und zu 17 Prozent ($n = 8$) interne Evaluationen. Bei den Auftraggebern der Schweiz liegen die Werte sogar bei 88 Prozent ($n = 45$; $N = 51$) interner und 12 Prozent ($n = 6$) externer Evaluationen. Der geschätzte Anteil an Evaluationsaufträgen, die direkt vergeben werden, liegt in den USA bei durchschnittlich 50 Prozent (SD = 38.2, $N = 42$), wobei 55 Prozent ($n = 23$) angeben, dass sie bis zur Hälfte der Evaluationen direkt vergeben und 45 Prozent der Befragten ($n = 19$) diesen Anteil zwischen 65 bis 100 Prozent einordnen. Für die Schweiz haben 53 Prozent ($n = 18$, $N = 34$) bis zu der Hälfte der Evaluationen direkt vergeben und die restlichen 47 Prozent ($n = 16$) der Auftraggeber haben zwei Drittel bis alle Evaluationen direkt vergeben. Demgegenüber zeigen sich Unterschiede in der Zusammensetzung der Auftraggeber bezüglich des Sektors, in dem sie tätig sind. Während die Auftraggeber der Schweiz fast ausschliesslich (92%, $n = 48$; $N = 52$) aus dem öffentlichen Sektor stammen, dominiert in den USA zwar

der öffentliche Sektor mit 40 Prozent (n = 21; N = 52) doch sind auch Auftraggeber aus anderen Sektoren vertreten (siehe Abbildung 4).

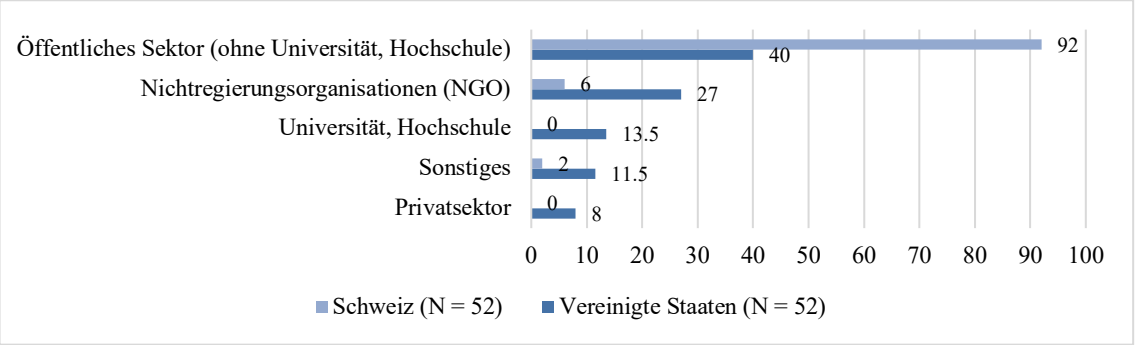

Abbildung 4: Sektor der Auftraggeber im Ländervergleich (eigene Darstellung)
Anmerkung: Die Zahlen repräsentieren Prozentwerte; Daten für die Schweiz stammen aus Pleger und Hadorn (2018, S. 6); Antworten auf die offene Frage wurden für beide Länder den fünf Kategorien zugeordnet (N = 52).

Darunter arbeiten 27 Prozent (n = 14) bei einer NGO, 13.5 Prozent (n = 7) bei einer Universität oder Hochschule, 11.5 Prozent (n = 6) in einem sonstigen Bereich und 8 Prozent (n = 4) im Privatsektor. Die Kategorie „Sonstiges" umfasst für die Schweiz eine selbstständig arbeitende Person, wobei für die USA eine Person bei der zwischenstaatlichen Organisation United Nations (UN) arbeitet sowie fünf weitere Personen im NPO-Sektor insb. im Bereich der Philanthropie und Stiftungen tätig sind. Aus Gründen der Vergleichbarkeit zur Studie in der Schweiz wurde die „NGO-Kategorie" nicht zur „NPO-Kategorie" erweitert.

Wird der Sektor der Evaluierenden im Ländervergleich (siehe Abbildung 5) angeschaut, zeigen sich auch hier Unterschiede. Die überwiegenden Vertragspartner der Auftraggeber beider Länder sind Evaluierende aus dem Privatsektor (USA: 45%, N = 49; CH: 60%, N = 50), gefolgt von jenen aus Universitäten oder Hochschulen (USA: 25%, CH: 56%).

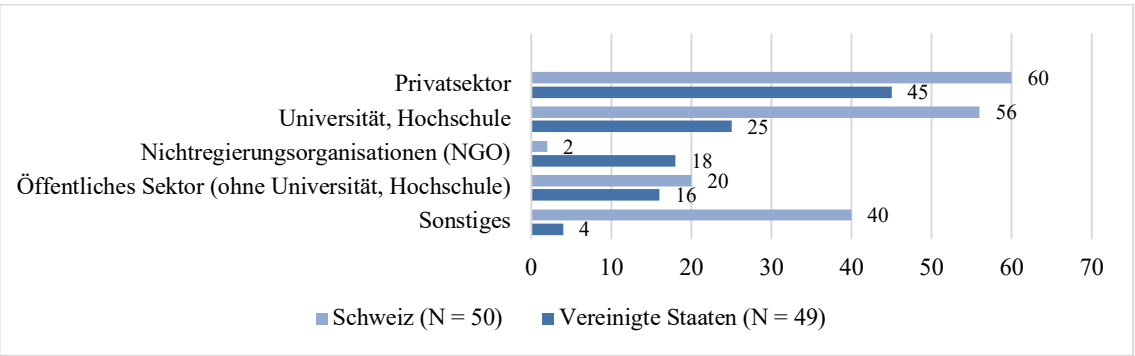

Abbildung 5: Sektor der Evaluierenden im Ländervergleich (eigene Darstellung)
Anmerkung: Die Zahlen repräsentieren Prozentwerte; Daten für die Schweiz stammen aus Pleger und Hadorn (2018, S. 7); Mehrfachantworten waren für die Schweiz möglich, wobei für die USA Antworten auf die offene Frage den fünf Kategorien zugeordnet wurden. Die Möglichkeit der Mehrfachantworten kann mit den tendenziell höheren Werten der Schweiz zusammenhängen. Die Kategorie „Sonstiges" enthält für die Schweiz 36% Selbstständige.

Weitere wichtige Vertragspartner in den USA arbeiten mit 18 Prozent (n = 9) als Evaluierende in NGO, gefolgt vom öffentlichen Sektor mit 16 Prozent (n = 8). Die Evaluierende in der Schweiz arbeiten mit 36% als Selbstständige (n = 18) (hier inkludiert in Sonstiges), wobei weitere 20 Prozent (n = 10) im öffentlichen Sektor und 2 Prozent (n = 1) in einer NGO tätig sind.

Welche Erwartungen Auftraggeber aus den USA und der Schweiz gegenüber der Unabhängigkeit von Evaluationen haben resp. wie sie Evaluationsstandards wahrnehmen wird mittels Tabelle 12 dargestellt.

Tabelle 12: Übersicht der Resultate zur Erwartung an die Unabhängigkeit im Ländervergleich

Ein guter Evaluierender …	Auftraggeber USA			Auftraggeber Schweiz		
	M	SD	*N*	M	SD	N
… ist primär festgelegten Evaluationsstandards verpflichtet	5.2	1.2	38	5.0	1.1	42
… präsentiert nur die von ihm tatsächlich ermittelten Ergebnisse	4.7	1.4	37	5.3	0.9	42
… lässt sich in seiner methodischen Vorgehensweise nicht beeinflussen	4.6	1.4	37	4.6	1.1	41
… legt schonungslos bei der Evaluation ermittelte Schwachstellen offen	4.5	1.5	39	5.0	1.2	44
… orientiert sich methodisch an den Interessen der Auftraggeber*	4.3	1.5	36	-	-	-
… hat in erster Linie eine moralische Verantwortung gegenüber den Stakeholdern	3.9	1.8	36	2.9	1.5	39
… richtet seine Evaluationsergebnisse nach den Erwartungen und Bedürfnissen der Auftraggeber	2.9	2.1	39	3.6	1.7	43
… schwächt negative Evaluationsergebnisse gegenüber Auftraggebern ab, damit sie eher akzeptiert werden*	1.7	1.1	39	-	-	-

Anmerkung: Skala von 0 (= Stimme überhaupt nicht zu) bis 5 (= Stimme voll zu). Zur Vergleichbarkeit wurden die Werte von 0 bis 5 an die Werte 1 bis 6 (siehe Pleger & Hadorn, 2018) angepasst. Daten für die Schweiz stammen aus Pleger und Hadorn (2018, S. 8). N = Gesamtzahl aller valider Fälle. *Da die beiden Items im Rahmen der vorliegenden Studie ergänzt wurden, liegen für die Schweiz diesbezüglich keine Daten vor.

(Quelle: eigene Darstellung)

Für den ungefähren Vergleich beider Länder in Bezug auf ihre durchschnittliche Erwartung an die Unabhängigkeit von Evaluationen wurde für jedes Land der Gesamtdurchschnitt der Mittelwerte der aufgeführten Items berechnet. So liegt die durchschnittliche Erwartung an die Unabhängigkeit von Evaluationen bei den Auftraggebern in den USA bei einem relativ tieferen Mittelwert von 3.975 und in der Schweiz bei einem Mittelwert von 4.4, wobei beide Werte einer hohen Erwartung entsprechen. Die Auftraggeber beider Länder erwarten insb. von Evaluierenden, dass sie sich primär festgelegten Evaluationsstandards verpflichten (USA: M = 5.2; CH: M = 5.0). Sowohl von den US-Auftraggebern (M = 4.7, SD = 1.4, N = 37) als auch den Schweizer Auftraggebern (M = 5.3, SD = .9, N = 42) wird erwartet, dass Evaluierende nur die tatsächlich ermittelten Ergebnisse präsentieren. Relativ hohe Erwartungen haben die Auftraggeber beider Länder auch dahingehend, dass sich Evaluierende nicht in der methodischen Vorgehensweise beeinflussen lassen (USA: M = 4.6, N = 37; CH: M = 4.6, N = 41) und schonungslos bei der Evaluation

ermittelte Schwachstellen offenlegen (USA: *M* = 4.5*, N* = 39*;* CH: *M* = 5.0, *N* = 44). Die US-Auftraggeber erwarten zudem mit einem durchschnittlichen Wert von 4.3 (SD = 1.5, *N* = 36), dass sich Evaluierende methodisch an den Interessen der Auftraggeber orientieren. Die Erwartung, dass Evaluierende eine moralische Verantwortung gegenüber Stakeholdern haben, wird stärker von den US-Auftraggebern (*M* = 3.9, SD = 1.8, *N* = 36) eingenommen, wobei dies für die Schweizer Auftraggeber weniger zutrifft (*M* = 2.9, SD = 1.5, *N* = 39). Dafür erwarten letztere vergleichsweise stärker, dass Evaluierende die Resultate nach den Erwartungen und Bedürfnissen der Auftraggeber ausrichten (*M* = 3.6, SD = 1.7, *N* = 43). Die Erwartung gegenüber Evaluierenden, dass sie negative Ergebnisse abschwächen, damit sie eher akzeptiert werden, wird beinahe nicht erwartet (*M* = 1.7, SD = 1.1, *N* = 39).

In Bezug auf die Evaluationsstandards werden deren Kenntnis, Vertrautheit und Wichtigkeit zwischen den Auftraggebern in den USA und der Schweiz vergleichend dargestellt (siehe Tabelle 13).

Tabelle 13: Übersicht der Resultate zur Standardkenntnis, Vertrautheit und Standardwichtigkeit im Ländervergleich

	Auftraggeber USA		Auftraggeber Schweiz	
	Prozent	N	Prozent	N
Standardkenntnis[a]				
Ja	30.0	9	68.0	18
Nein	70.0	21	32.0	8
		30		**26**
Vertrautheit[b]				
Überhaupt nicht bis eher schlecht vertraut	22.3	2	37.0	10
Eher gut bis sehr gut vertraut	77.8	7	63.0	16
		9		**26**
Standardwichtigkeit[c]				
Überhaupt nicht wichtig	0.0	0	0.0	0
Eher unwichtig	11.1	1	15.0	4
Eher wichtig	77.8	7	46.0	12
Sehr wichtig	11.1	1	39.0	10
		9		**26**

Anmerkung: Daten für die Schweiz stammen aus Pleger und Hadorn (2018, S. 12); a. Die Variable *Standardkenntnis* wurde mit der Frage gemessen: Kennen Sie die nationalen Evaluationsstandards „Program Evaluation Standards" des JCSEE (Joint Committee on Standards on Educational Evaluation"? b. Die Variable *Vertrautheit* wurde mit der Frage gemessen: Wie gut sind Sie mit den nationalen Evaluationsstandards (Program Evaluation Standards) vertraut? c. Die Variable *Standardwichtigkeit* wurde mit der Frage gemessen: Wie wichtig finden Sie die nationalen Evaluationsstandards (Program Evaluation Standards) für den Evaluationsprozess?

(Quelle: eigene Darstellung)

Während 68 Prozent der Schweizer Respondenten (*n* = 18) die nationalen Evaluationsstandards kennen, zeigt sich das Gegenteil für die Auftraggeber der USA, wobei nur 30 Prozent (*n* = 9) die nationalen Evaluationsstandards kennen. Unter denjenigen Auftraggebern, welche die Standards kennen, sind in den USA 78 Prozent (*n* = 7, *N* = 9) und in

der Schweiz 63 Prozent (n = 16, N = 26) eher gut bis sehr gut damit vertraut. Die Einstellungen in Bezug auf die Standardwichtigkeit divergierten jedoch zwischen den Auftraggebern der beiden Länder. 39 Prozent der Schweizer Auftraggeber (n = 10) bewerten die Standards als sehr wichtig, wobei dies in den USA nur bei einer Person (11%, N = 9) zutrifft. Insgesamt bewerten die Auftraggeber beider Länder die Standards als eher wichtig, wobei es in den USA 78 Prozent (n = 7) und in der Schweiz 46 Prozent (n = 10) sind. Kein Auftraggeber ordnet die Standards als überhaupt nicht wichtig ein.

4.4.2 Beziehungsverhältnis zwischen Auftraggebern und Evaluierenden im Ländervergleich

Das Beziehungsverhältnis zwischen Auftraggebern und Evaluierenden wird hinsichtlich unterschiedlicher wahrgenommener Aspekte von Auftraggebern aus den Vereinigten Staaten und der Schweiz verglichen. Erstens wird auf die Reaktion auf Änderungsvorschläge sowie die Unterstellung von Druckausübung und Einflussnahme eingegangen. Zweitens werden von den Auftraggebern wahrgenommene Ursachen für Schwierigkeiten der Zusammenarbeit mit Evaluierenden im Ländervergleich erörtert. Abschliessend werden die von den Aufraggebern vorgeschlagenen präventiven Massnahmen verglichen.

Gemäss den Angaben der Auftraggeber der USA und Schweiz wurde nach der Unterbreitung von Änderungsvorschlägen seitens der Auftraggeber mehrheitlich ein Kompromiss zwischen den Evaluierenden und den Auftraggebern gefunden (siehe Tabelle 14).

Tabelle 14: Übersicht der Resultate zur Reaktion und Unterstellung im Ländervergleich

	Auftraggeber USA		Auftraggeber Schweiz	
	Prozent	N	Prozent	N
Reaktion[a] Änderungsvorschläge				
Die Änderungen wurden vorgenommen	3.3	1	3.0	1
Die Änderungen wurden nicht vorgenommen	26.7	8	44.0	14
Es wurde ein Kompromiss gefunden	70.0	21	53.0	17
		30		**32**
Unterstellung[b] Druckausübung/Beeinflussung				
Nein, dies wurde mir noch nie unterstellt	93.8	30	92.0	36
Ja, dies wurde mir schon unterstellt	0.0	0	3.0	1
Weiss nicht	6.2	2	5.0	2
		32		**39**

Anmerkung: Daten für die Schweiz stammen aus Pleger und Hadorn (2018, S. 10-12)

a. Die Variable *Reaktion* wurde mit der Frage gemessen: Wie reagierten die von Ihnen beauftragten Evaluierenden überwiegend darauf, als Sie sie um Änderungen baten?

b. Die Variable *Unterstellung* wurde mit der Frage gemessen: Wurde Ihnen schon einmal von einem Evaluator unterstellt, ihn unter Druck gesetzt oder beeinflusst zu haben, Ergebnisse falsch oder ungenau darzustellen?

(Quelle: eigene Darstellung)

Dies gilt für 70 Prozent der US-Auftraggeber ($n = 21$, $N = 30$) und 53 Prozent der Schweizer Auftraggeber ($n = 17$, $N = 32$). Jeweils eine Person gab an, dass die Änderungen vorgenommen wurden, wobei dies bei 27 Prozent ($n = 8$) der Auftraggeber in den USA nicht der Fall war. In der Schweiz sind es sogar 44 Prozent der Auftraggeber ($n = 14$) bei denen die Änderungen nicht vorgenommen wurden. Wird die Unterstellung einer Druckausübung oder Beeinflussung zwischen den Auftraggebern der beiden Länder verglichen, zeigt sich ein klarer, übereinstimmender Trend, dass ihnen (fast) nie eine solche Unterstellung unterbreitet wurde. Beinahe allen Auftraggebern der USA (94%, $n = 30$) wurde weder eine Druckausübung noch eine Beeinflussung unterstellt, wobei 2 Personen (6%) ihre Antwort mit „Weiss nicht" eingeordnet haben ($N = 32$). 92 Prozent der Schweizer Auftraggeber ($n = 36$) wurde dies noch nie unterstellt und auch zwei Personen (5%) haben mit „Weiss nicht" geantwortet. Im Gegensatz zur USA hat ein Auftraggeber in der Schweiz (3%) die Frage nach der Unterstellung bejaht ($N = 39$).

Obwohl die Gegenüberstellung beider Länder nicht auf der gleichen Messung basiert, geben die Prozentwerte Anhaltspunkte darüber, welche Aspekte für die Schwierigkeiten in der Zusammenarbeit zwischen Auftraggebern und Evaluierenden verantwortlich sein können. Im Ländervergleich fällt auf, dass für beide Länder die gleiche Hauptursache für Schwierigkeiten genannt wird (siehe Tabelle 15).

Tabelle 15: Übersicht der Resultate zu den Schwierigkeiten in der Zusammenarbeit im Ländervergleich

Die Zusammenarbeit mit Evaluierenden ist generell schwierig, weil …	Auftraggeber USA		Auftraggeber Schweiz	
	Ja	N		N
… den Evaluierenden das Verständnis für die zu evaluierende Organisation fehlt	51.4 (18)	35		36
… das gegenseitige Verständnis zwischen den Evaluierenden und den Auftraggebern fehlt	45.7 (16)	35	42.0 (15)[a]	36
… die Unabhängigkeit von Evaluationen gewährleistet werden soll, ohne dass ich als Auftraggeber den Evaluierenden zu stark beeinflusse	26.5 (9)	34	19.0 (7)	36
… den Evaluierenden wichtige Fachkompetenzen fehlen	44.1 (15)	34	14.0 (5)	36
… persönliche Faktoren der Evaluierenden in die Evaluation einfliessen können (z. B. politische Einstellung, persönliche Präferenzen, etc.)	37.1 (13)	35	14.0 (5)	36
… die Evaluierenden zu wenig Ressourcen (z. B. Zeit) für ihre Arbeit haben	40.6 (13)	32	11.0 (4)	36
… die Evaluierenden nicht motiviert sind*	9.1 (3)	33	-	-

Anmerkung: Für die USA wurde die Variable *Schwierigkeiten* mit der Frage gemessen: Worin sehen Sie generell die grössten Schwierigkeiten in der Zusammenarbeit mit Evaluierenden? mit Nein (= 0) oder Ja (= 1). N = Gesamtzahl aller Fälle. Die Zahlen repräsentieren Prozentwerte, in Klammern wurden die jeweiligen Häufigkeiten angegeben; Daten für die Schweiz stammen aus Pleger und Hadorn (2018, S. 9). Dort wurde die Variable anhand einer offenen Frage erhoben, wobei die Antworten einzelnen Kategorien zugeordnet wurden. a. Die vorliegende Studie hat die ersten beiden Items eindimensional gemessen (vorher eine Kategorie). *Da dieses Item im Rahmen der vorliegenden Studie ergänzt wurde, liegen für die Schweiz keine Daten vor.

51 Prozent (n = 18; N = 35) resp. 46 Prozent der US-Auftraggeber (n = 16) haben das fehlende Verständnis für die zu evaluierende Organisation resp. das fehlende gegenseitige Verständnis zwischen den Evaluierenden und den Auftraggebern als Ursache für die schwierige Zusammenarbeit genannt. 42 Prozent der Schweizer Auftraggeber (n = 15; N = 36) nannten diese Ursache ebenfalls am häufigsten. Die Unabhängigkeit von Evaluationen und die Beeinflussung des Auftraggebers haben 19 Prozent der Schweizer Auftraggeber (n = 7) als zweithäufigste Ursache angegeben. Für die USA wurden vielmehr Ursachen wie das Fehlen von wichtigen Fachkompetenzen (44%, n = 15) sowie Ressourcen (41%, n = 13), aber auch persönliche Faktoren (37%, n = 13) als Einflussquellen für die schwierige Zusammenarbeit genannt. Dass Evaluierende nicht motiviert sind, wurde in den USA lediglich von 9 Prozent (n = 3) angegeben.

Für eine detaillierte Beschreibung zu den Angaben der präventiven Massnahmen seitens der Auftraggeber der USA wird auf Kapitel 4.3.2 verwiesen. Die Massnahmen sind im Ländervergleich in Abbildung 6 dargestellt.

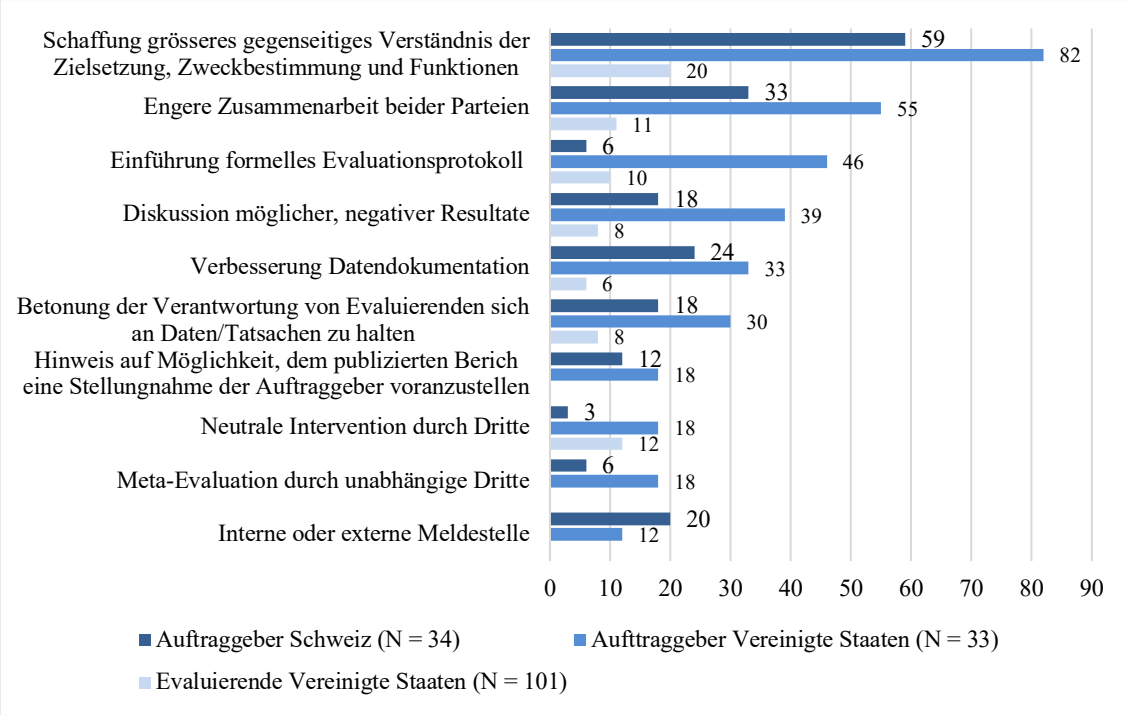

Abbildung 6: Präventive Massnahmen im Ländervergleich (eigene Darstellung)
Anmerkung: Die Zahlen repräsentieren Prozentwerte; Mehrfachantworten waren für die Auftraggeber als Respondenten möglich, nicht aber für die Evaluierenden; Daten für die Schweiz stammen aus Pleger und Hadorn (2018, S. 13); Daten für die Evaluierenden stammen aus (Morris & Clark, 2012, S. 62)

Die Auftraggeber beider Länder nennen die Schaffung eines grösseren gegenseitigen Verständnis (USA: 82%, CH: 59%) sowie eine engere Zusammenarbeit zwischen Auf-

traggebern und Evaluierenden (USA: 55%, CH: 33%) am häufigsten als präventive Massnahmen. Für die Schweizer Auftraggeber wird danach die Verbesserung der Datendokumentation (24%) genannt, wobei bei den US-Auftraggebern die Einführung eines formellen Evaluationsprotokolls (46%) drittrangig genannt wird. Diese Massnahme wird für die Schweiz zu lediglich 6 Prozent genannt. Die Diskussion möglicher, negativer Resultate (USA: 39%, CH: 18%), die Betonung der Verantwortung der Evaluierenden (USA: 30%, CH: 18%) sowie der Hinweis auf eine mögliche Stellungnahme (USA: 18%, CH: 12%) ordnen die Länder in dieser Reihenfolge als weitere wichtige Massnahmen ein. Für die Schweizer Auftraggeber kommen den beiden Massnahmen der neutralen Intervention von Dritten (3%) und der Meta-Evaluation durch unabhängige Dritte (6%) relativ weniger Bedeutung zu. Jeweils 18 Prozent der US-Auftraggeber nennen diese Massnahmen. Eine interne oder externe Meldestelle nennen 20 Prozent der Schweizer Auftraggeber und 12 Prozent der US-Auftraggeber. Werden die Ergebnisse zusätzlich mit den Evaluierenden in den Vereinigten Staaten verglichen, fällt auf, dass die meistgenannte präventive Massnahme des grösseren gemeinsamen Verständnis mit denjenigen der Auftraggeber der Schweiz und den USA übereinstimmen. Danach nennen die Evaluierenden die neutrale Intervention durch Dritte relativ häufig mit 12 Prozent (n = 12), was als Massnahme v.a. für die Auftraggeber der Schweiz vergleichsweise unbedeutend ist. Gegensätzlich wird auch die Massnahme der verbesserten Datendokumentation eingeordnet, wobei sie für die Evaluierenden mit 6 Prozent (n = 6) vergleichsweise unbedeutend scheint. Die enge Zusammenarbeit, die Einführung eines formellen Evaluationsprotokolls sowie die Diskussion möglicher, negativer Resultate stimmt von der Rangordnung her mit derjenigen der US-Auftraggeber überein. Im nächsten letzten Abschnitt werden abschliessend die Gütekriterien erörtert.

4.5 Gütekriterien

Nachfolgend werden die Gütekriterien der Objektivität, Reliabilität und Validität auf das Messinstrument des Fragebogens mit den darin enthaltenen Skalen angewendet.

Die Ergebnisse fallen während der Erhebungsdurchführung vollständig unabhängig von der Verhaltensweise der Autorin aus, was die Durchführungsobjektivität erfüllt. Auch die Auswertungsobjektivität ist aufgrund der standardisierten Befragung und der durch das Codebuch gewährleisteten intersubjektiven Nachvollziehbarkeit gegeben. Die Interpretationsobjektivität ist für die geschlossenen Fragen ganzheitlich erfüllt, für die offenen Fra-

gen bei der Zuordnung in jeweilige Kategorien jedoch abhängig von der Untersuchungsperson. Die Objektivität ist jedoch insgesamt erfüllt (Sedlmeier & Renkewitz, 2018, S. 80).

Generell fällt die Reliabilität der Messinstrumente gering bis mittel aus. Die Ergebnisse der Befragung können nämlich aufgrund unkontrollierter und unsystematischer Einflüsse schwanken, da das Interesse, die Motivation, aber auch die Müdigkeit von Befragten die Antworten beeinflussen können. Weitere Einflüsse können bei wiederholter Erhebung auch bei veränderten Untersuchungssituationen auftreten wie bspw. einer anderen Raumtemperatur oder Tageszeit während der Befragungsteilnahme. Bei wiederholter Messung könnten verschiedene Items von den Respondenten unterschiedlich aufgefasst werden, was zu anderen Antworten führt (Ebd., 2018, S. 81). Diese Einflüsse sind unvermeidbar, bewegen sich aber in einem üblichen und für die vorliegende Forschung vertretbaren Rahmen. Die Homogenitätsgrade der Skalen liegen grundsätzlich in einem akzeptablen Bereich, wobei einige Kernkonstrukte hohe und andere eher tiefe Cronbach's Alpha Werte aufweisen. Insgesamt kann gesagt werden, dass der Varianzanteil in den Testwerten, der auf die Varianz der wahren Werte zurückgeht, mittelmässig ausgeprägt ist (siehe Kapitel 3.3) (Krüger et al., 2012, S. 48).

Die Inhaltsvalidität des Messinstruments ist insofern gegeben, dass die Operationalisierung zentraler Konstrukte auf dem BUSD-Modell basiert oder auf Skalen zurückgegriffen wurde, die bereits in vergangener Forschung erfolgreich verwendet wurden. In beiden Fällen haben Experten der Evaluationsforschung sichergestellt, dass das Modell resp. die Skalen die wesentlichen Aspekte des für das zu messende Merkmal erfassen (Ebd., 2018, S. 86). Die durchgeführten Hypothesentests liefern nur partielle Erkenntnisse bezüglich der Konstruktvalidität und zwar zeigen einige Befunde, dass gewisse Konstrukte mit anderen Variablen in theoretisch begründbaren Zusammenhang stehen. Für genauere Aussagen über das Messinstrument und dessen Konstruktvalidität – aber auch der Kriteriumsvalidität – müssen die Konstrukte in weiteren Untersuchungen überprüft werden (Diekmann, 2013, S. 258–259). Insgesamt wurde das Ziel der Konstruktion eines objektiven, reliablen und validen Messinstruments eingehalten (Ebd., 2013, S. 261). Im nächsten Kapitel werden die Resultate systematisch diskutiert, wobei die Forschungsfragen schrittweise beantwortet werden.

5 Diskussion

In diesem Kapitel werden die Resultate, gegliedert nach den Forschungsfragen, zusammengefasst und vor dem Hintergrund der Principal-Agent-Theorie und dem aktuellen Forschungsstand diskutiert. Gleichzeitig wird die Hauptforschungsfrage *wie die Unabhängigkeit von Evaluationen in den USA beurteilt wird* und *welche Rolle die Eigenschaften der Auftraggeber und ihre Beziehung zu Evaluierenden in der Beeinflussung des Evaluationsprozesses spielen* anhand der untergeordneten Forschungsfragen schrittweise beantwortet.

85 Prozent der Befragten haben mindestens schon einmal destruktiven Einfluss auf Evaluierende ausgeübt ($N = 27$). Besonders häufig wurde die Umformulierung einzelner wertender Sätze im Evaluationsbericht sowie die Modifikation des Fragebogens oder Reports gefordert oder es wurde darauf hingewiesen, wie andere Evaluierende vorgegangen wären. Diese Befunde stimmen mit den Erkenntnissen überein, dass Auftraggeber als die grösste Einflussquelle identifiziert wurden, wobei gleichzeitig die Evaluationen nicht als unabhängig wahrgenommen wurden (Pleger et al., 2016, S. 1). 42 Prozent der US-Evaluierenden ($N = 905$) haben bereits eine Druckausübung erlebt, die Resultate falsch darzustellen, wovon dies 70 Prozent mehrmals erlebt haben (Morris & Clark, 2012, S. 61). Dabei erstaunt die Divergenz zwischen der relativ stark ausgeprägten destruktiven Einflussnahme unter den US-Auftraggebern und den relativ tiefer ausfallenden, dokumentierten Druckausübungen der betroffenen Evaluierenden der USA. Dieser Vergleich ist jedoch mit Vorsicht zu interpretieren, da die zugrundeliegende Methodik und Fallzahl unterschiedlich sind. Auch überrascht, dass doch 30 Prozent der Befragten mindestens einmal gefordert haben gewisse Teile wegzulassen. Ebenso haben jeweils gut ein Viertel mindestens einmal vorgeschlagen, die Resultate positiver darzustellen sowie eine Person hat den Evaluierenden im Vorhinein mitgeteilt, welche Ergebnisse erwartet werden. Bei Berücksichtigung der Antworten auf die offenen Fragen zeigt sich, dass viele Änderungsvorschläge der Auftraggeber auch konstruktiver Art sind. Dies wird durch die Antwort treffend verdeutlicht, „dass Änderungsempfehlungen gemacht werden, um die Ergebnisse für ein breites Publikum besser verständlich zu machen und nicht, um Ergebnisse oder Konsequenzen von Befunden zu verändern“. Trotzdem enthielten andere Antworten wiederum deutliche Hinweise einer destruktiven Einflussnahme. Eine Person rechtfertigte die Modifikation des Fragebogens insofern, dass ihre Evaluierende nutzenfokussiert seien und daher Anpassungen an den Kontext und die Umwelt durchaus angemessen

seien. Ob diese Äusserung tatsächlich als destruktiv oder konstruktiv eingeordnet werden kann, kann vorliegend nicht beurteilt werden. Auch wurde erwähnt, dass Änderungen zu sachdienlicheren Informationen führen sollten oder aufgrund vertraglicher Bestimmungen durchgeführt wurden. Die *allgemeine destruktive Beeinflussungsart* zeigt ein restriktiveres Bild in Bezug auf die generelle destruktive Einflussnahme der Auftraggeber, was jedoch mit der tiefen internen Konsistenz (Cronbach's Alpha: .523) dieser kurzen Skala zusammenhängen kann (Hemmerich, 2019). Während bei 37 Prozent der Auftraggeber mindestens einmal vorgekommen ist, dass sich der Evaluationsprozess aufgrund von Änderungsvorschlägen verzögert hat (*Undermining)*, haben 34 Prozent mindestens einmal Evaluierende aufgefordert gewisse Evaluationsresultate zu kürzen (*Distortion)*. Das häufigere Auftreten der indirekten Beeinflussungsform *Undermining* als *Distortion* deckt sich mit dem Befund von Pleger und Sager (2018, S. 170). Die *Beeinflussungsintention* unterstreicht, dass die destruktive Beeinflussungsform bei allen Respondenten in verschiedenen Intensitäten ausgeprägt ist. Trotz der kleinen Samplegrösse bekräftigen die Befunde die Verbreitung und Relevanz der destruktiven Beeinflussungsart. Aufgrund der Übereinstimmung mit den Befunden der Schweizer Studie wird zusätzlich deren Validität bestätigt, wobei von einem Phänomen in Ländern mit einer ausgeprägten Evaluationskultur ausgegangen werden kann.

Generell gestaltet sich das Beziehungsverhältnis zwischen Auftraggebern und Evaluierenden gemäss den Angaben der US-Auftraggeber als mittelmässig konfliktgeprägt. Davon hatten jeweils rund die Hälfte jemals einen Konflikt mit Evaluierenden aufgrund deren Arbeit oder Arbeitsweise, wobei dies bei 44 Prozent der Schweizer Auftraggeber zutrifft. Diese Befunde sind deckungsgleich, was zugleich deren Validität im Ländervergleich unterstreicht. Obwohl aufgrund der fehlenden statistischen Signifikanz keine sicheren Aussagen über die abgelehnte H1 gemacht werden können, liefern die Befunde Hinweise darauf, dass je stärker Auftraggeber ein konfliktgeprägtes Verhältnis mit Evaluierenden wahrnehmen, desto stärker die destruktive Beeinflussungsart ausgeprägt ist. Statistisch signifikante Befunde würde bedeuten, dass die dem Interessenkonflikt zugrundeliegende divergierende Interessen der Auftraggeber und Evaluierenden mit kontrollierenden Beeinflussungsversuchen der Auftraggeber zusammenhängen. Demgegenüber wird der erste Teil der H4 aufgrund der signifikanten Testwerte vorläufig bestätigt, wobei angenommen werden kann, dass je stärker die Auftraggeber ein konfliktgeprägtes Verhältnis mit den Evaluierenden wahrnehmen, desto häufiger die Auftraggeber Anreizsysteme einsetzen. Damit bestätigt sich die dieser Hypothese zugrundeliegende Annahme

der PAT vorläufig, dass Auftraggeber die divergierenden Interessen, die zu einem Interessenkonflikt führen, durch Anreize oder Sanktionen gegenüber Evaluierenden einzuschränken versuchen (Jensen & Meckling, 1976, S. 308). Durch das Anreizsystem versuchen die Auftraggeber die Interessen der Evaluierenden an ihre eigenen anzugleichen, um damit das Risiko eines möglichen eigennützigen Verhaltens der Evaluierenden zu minimieren (Van Slyke, 2006). Wird den Evaluierenden mitgeteilt, dass ihnen der Auftrag entzogen werden könnte, kann abgeleitet angenommen werden, dass sich dadurch die Interessen der Evaluierenden ändern. Dabei kann eine Angleichung ihrer Interessen an diejenigen der Auftraggeber angenommen werden, die einerseits zur Zufriedenheit der Auftraggeber mit dem Auftrag und andererseits zu möglichen Anpassungen der Evaluationsresultate zugunsten der Auftraggeber führen. Die Forschungsfrage *inwiefern ein von den Auftraggebern als konfliktgeprägt wahrgenommenes Evaluationsverhältnis mit ihrer Einflussnahme auf den Evaluationsprozess zusammenhängt* kann damit lediglich für den signifikanten Zusammenhang sicher beantwortet werden, dass ein als konfliktgeprägt wahrgenommenes Verhältnis positiv mit dem Anreizsystem zusammenhängt. Insgesamt deuten aber Hinweise darauf, dass ein als konfliktgeprägt wahrgenommenes Evaluationsverhältnis positiv mit der Einflussnahme der Auftraggeber auf den Evaluationsprozess zusammenhängt, wobei sich die Einflussnahme in der destruktiven Beeinflussungsart, dem Anreizsystem und dem direkten Einfluss äussert. *Wie divergierende Interessen seitens der Auftraggeber ausgeglichen werden,* kann anhand des Anreizsystems erklärt werden, das die Auftraggeber während dem Evaluationsprozess einsetzen. So verwenden Auftraggeber v. a. die Massnahme der Mitteilung gegenüber den Evaluierenden sie nicht mehr für zukünftige Aufträge zu berücksichtigen, die Erklärung, dass ihnen der Auftrag entzogen werden könnte oder das Nahelegen von klaren Anreizen für die Ergebnisänderung (z. B. anhand von Nachfolgeaufträgen), um die divergierenden Interessen auszugleichen. Die erste Massnahme äussert sich dabei in einer indirekten, die zweite und dritte Massnahme in einer direkten, konkreten Einflussnahme.

Die Unzufriedenheit mit vergangenen Evaluationen fällt im Ländervergleich sehr ähnlich aus, wobei 57 Prozent der Schweizer Auftraggeber und 53 Prozent der US-Auftraggeber ihre Unzufriedenheit angaben. Analog zur Schweizer Studie stellte sich die ungenügende Evaluationsqualität ebenso als Hauptursache der Unzufriedenheit heraus, gefolgt von den nichterfüllten Erwartungen. Im Rahmen der offenen Frage gab eine Person als Unzufriedenheitsgrund an, dass „der Evaluationsbericht die Stärken und Erfolge des Projekts nicht

so sehr betont“. Diese Aussage unterstreicht einen möglichen in der H2 postulierten positiven Zusammenhang der *Unzufriedenheit* und der *destruktiven Beeinflussungsart*. Als weiterer Unzufriedenheitsgrund wurde erwähnt, dass „der Evaluierende die Ergebnisse des Berichts, die während der mündlichen Besprechung erläutert wurden, gedämpft hatte“. Dieser Befund kann insofern besorgniserregend sein, dass der Evaluierende tatsächlich vorhandene positive Effekte im Bericht gedämpft hatte. Gegebenenfalls spielen aber hier die von Morris (2007, S. 413) diskutierte Annahme eine Rolle, dass Evaluierende bestrebt sind die von Auftraggebern als erwünscht wahrgenommenen Ergebnisse zu berichten. Für den beschriebenen Fall gibt es zwei mögliche Szenarien und zwar, erstens dass sich der Evaluierende nach einer übertriebenen mündlichen Äusserung doch auf die tatsächlichen Resultate berufen hat oder die tatsächlichen, mündlich geäusserten Befunde im Nachhinein abgeschwächt hat. Unabhängig davon was zutrifft, signalisiert dieser Befund, dass nicht ausgeschlossen werden kann, dass auch Evaluierende selber Einfluss auf die Evaluationsresultate ausüben und Evaluationsresultate falsch darstellen können. Die Unzufriedenheit der Auftraggeber hinsichtlich unterschiedlicher Aspekte der Evaluation ist ambivalent ausgeprägt. Jeweils 32 Prozent der Auftraggeber waren insgesamt ziemlich zufrieden oder ziemlich unzufrieden. Besonders unzufrieden waren die US-Auftraggeber mit den Aspekten der methodischen Vorgehensweise, der Evaluations- und Methodenkompetenz der Evaluierenden sowie der Evaluationsqualität. Auffällig dabei ist, dass diese Unzufriedenheitsgründe mit den Konfliktgründen der US-Auftraggeber, aber auch derjenigen der Schweizer Auftraggeber korrespondieren. Der meistgenannte Konfliktgrund unter den US-Auftraggebern liegt im Fehlen wichtiger Kompetenzen seitens der Evaluierenden sowie der mangelhaften Qualität der Evaluationsresultate. Diese Befunde erstaunen nicht, denn die Auftraggeber delegieren den Evaluierenden den Evaluationsauftrag genau aufgrund ihrer spezialisierten Evaluationskompetenz, die sich in einem Informationsvorteils bezüglich dieser sachlichen Aufgabenbearbeitung niederschlägt. Nehmen die Auftraggeber „ex post“ eine gewisse Inkompetenz wahr, scheint es naheliegend, dass sich diese Wahrnehmung in einer Unzufriedenheit äussert oder gar in einen Konflikt resultiert und mit kontrollierenden Beeinflussungsversuchen zusammenhängt. Die Zusammenhangsmasse liefern Hinweise auf einen positiven, statistisch signifikanten Zusammenhang zwischen der *Unzufriedenheit* und der *destruktiven Beeinflussungsart*, sodass die H2 vorläufig bestätigt wird. Die Befunde weisen somit darauf hin, dass sich der Interessenkonflikt zwischen den Auftraggebern und Evaluierenden durch eine gewisse Unzufriedenheit mit verschiedenen Aspekten einer Evaluation äussert, die

zugleich positiv mit kontrollierenden, destruktiven Beeinflussungsversuchen der Auftraggeber zusammenhängt.

Für den ähnlichen, angenommenen positiven Zusammenhang zwischen den *Schwierigkeiten* und der *destruktiven Beeinflussungsart* zeigt sich kein statistisch signifikanter Zusammenhang, weshalb H3 abgelehnt wird. Im Ländervergleich stimmen die wahrgenommenen Hauptursachen für die Schwierigkeiten miteinander überein, wobei das fehlende Verständnis für die zu evaluierende Organisation sowie das fehlende gegenseitige Verständnis zwischen den Evaluierenden und Auftraggebern dominieren. Insofern hätte es nicht erstaunt, dass wie bei H3 angenommen, die Auftraggeber kontrollierenden Einfluss auf den Evaluationsprozess ausüben, wenn sie sich falsch verstanden fühlen oder wahrnehmen, dass der Evaluationsgegenstand missverstanden wurde. Für die USA wurden v. a. Ursachen wie das Fehlen von wichtigen Fachkompetenzen sowie Ressourcen, aber auch persönliche Faktoren als Einflussquellen für die schwierige Zusammenarbeit genannt. Eine Person gab an, dass „sich Evaluierende auf das aktuelle Paradigma beschränken und nicht verstehen wie in einem Umfeld von schnellem Wachstum und Innovation gedacht und evaluiert werden soll“. Dieser Kommentar stimmt mit Picciotto's (2019, S. 95) Ausführungen bezüglich einer notwendigen Kompetenzerweiterung von Evaluierenden überein. Damit in Zeiten des schnellen Wandels die zunehmende Lücke spezialisierter Evaluationen geschlossen werden kann, müssten Evaluierende ihre Instrumente erweitern und ihre Kompetenzen verbessern. Um den Erwartungen der Auftraggeber gerecht zu werden, müssten Evaluierende innovative Ansätze verfolgen und gleichzeitig in Managementsysteme und soziale Prozesse eingebettet werden. Die Forschungsfrage *wie die negative Wahrnehmung der Auftraggeber gegenüber Evaluierenden mit ihrer Einflussnahme auf den Evaluationsprozess zusammenhängt,* kann für den signifikanten Zusammenhang sicher beantwortet werden, dass die *Unzufriedenheit* positiv mit der *destruktiven Beeinflussungsart* der Auftraggeber zusammenhängt. Je stärker die Auftraggeber mit einer in Auftrag gegebenen Evaluation unzufrieden sind, desto stärker ist auch ihre destruktive Beeinflussungsart ausgeprägt. Für den postulierten positiven Zusammenhang zwischen den wahrgenommenen *Schwierigkeiten* in der Zusammenarbeit mit Evaluierenden gibt es weder eindeutige Hinweise bezüglich der Zusammenhangsrichtung, noch existiert ein signifikanter Zusammenhang. Somit kann keine Aussage über den in H3 angenommenen Zusammenhang gemacht werden.

Die *Erwartung an die Unabhängigkeit* von Evaluationen liegt im Ländervergleich auf einem eher hohen Niveau. Insgesamt haben die Auftraggeber beider Länder hohe Erwartungen an die Unabhängigkeit von Evaluationen. Sowohl US-Auftraggeber als auch Schweizer Auftraggeber erwarten, dass sich Evaluierende primär festgelegten Evaluationsstandards verpflichten, sich nicht in der methodischen Vorgehensweise beeinflussen lassen, nur die tatsächlich ermittelten Ergebnisse präsentieren und schonungslos bei der Evaluation ermittelte Schwachstellen offenlegen. Der Befund, dass lediglich 30 Prozent der US-Auftraggeber die nationalen Evaluationsstandards kennen, überrascht und ist alarmierend zugleich. Gerade in Anbetracht der eher hohen Erwartungen an die Unabhängigkeit von Evaluationen scheint der Befund widersprüchlich. Bei den Schweizer Auftraggebern kennen immerhin 68 Prozent der Befragten die Standards. Für beide Länder wird somit eine divergierende Rollenwahrnehmung beobachtet, die sich in der hohen Erwartung äussert, dass sich Evaluierende primär festgelegten Evaluationsstandards verpflichten. Gleichzeitig erheben die Auftraggeber diesen Anspruch aber nicht für sich selbst, was sich ebenso in der tiefen Bekanntheit der Evaluationsstandards zeigt (Pleger & Hadorn, 2018, S. 14). Hinsichtlich der *Vertrautheit* sind die US-Auftraggeber besser mit den Evaluationsstandards vertraut als die Schweizer Auftraggeber, wobei jedoch die geringe Fallzahl der vorliegenden Studie zu beachten und der Befund mit Vorsicht zu interpretieren ist. Unter denjenigen Auftraggebern, welche die Standards kennen, sind in der Schweiz 78 Prozent und in den USA 63 Prozent eher gut bis sehr gut damit vertraut. Dennoch bleibt die Frage unbeantwortet: wieso sind die nationalen „Program Evaluation Standards“ den US-Auftraggebern so wenig bekannt? Die berechneten Zusammenhangsmasse zur Überprüfung der H5 liefern Hinweise darauf, dass die *Vertrautheit* positiv mit der *Erwartung an die Unabhängigkeit* zusammenhängt. Die Forschungsfrage, *ob es einen Zusammenhang zwischen der Vertrautheit von Auftraggeber mit Evaluationsstandards und deren Erwartung an die Unabhängigkeit von Evaluationen gibt*, kann aufgrund der fehlenden statistischen Signifikanz der Zusammenhangsmasse nicht sicher beantwortet werden. Trotzdem gibt es Hinweise auf die Richtung eines möglichen Zusammenhangs, dass je mehr Auftraggeber mit den Evaluationsstandards vertraut sind, desto höher die Erwartungen an die Unabhängigkeit von Evaluationen ausgeprägt sind. Je geringer also das Informationsdefizit der Auftraggeber hinsichtlich der nationalen „Program Evaluation Standards“ ausfällt, desto höher sind die Erwartungen an die Unabhängigkeit von Evaluationen ausgeprägt. Statistisch signifikante Befunde würden somit die Wichtigkeit, dass die Evaluationsstandards unter den US-Auftraggebern bekannt und entsprechend

vertraut sind unterstreichen, damit im Umkehrschluss die Unabhängigkeit von Evaluationen gewährleistet oder gar verbessert werden könnte. Durch eine stärkere Vertrautheit mit Evaluationsstandards seitens der Auftraggeber würde sich die damit verbundene Informationsasymmetrie zu den Evaluierenden minimieren, indem beide Akteure ein ähnliches Evaluationsverständnis teilen und sich die divergierenden Rollenwahrnehmungen beider Akteure angleichen würden. Vor dem Hintergrund der tiefen Bekanntheit dieser Standards wird jedoch auf die notwendige Reduzierung dieses alarmierenden Informationsdefizits seitens der Auftraggeber plädiert.

Bei der *konstruktiven Beeinflussungsart* fällt auf, dass alle Auftraggeber mindestens mehr als einmal konstruktiven Einfluss auf den Evaluationsprozess ausgeübt haben, wobei sogar 61 Prozent der Befragten angegeben haben, dies schon oft gemacht zu haben. Auch bei der *konstruktiven Beeinflussungsart Allgemein* haben 90 Prozent der US-Auftraggeber mehr als einmal konstruktiven Einfluss ausgeübt. Das Ergebnis verdeutlicht, dass die Einflussnahme von Auftraggebern eben nicht nur negativer, destruktiver, sondern auch positiver, konstruktiver Art sein kann. In der Zusammenarbeit mit Evaluierenden unterstützen die Auftraggeber die Evaluierenden, indem sie relevante Informationen frühzeitig liefern und konstruktive Dialoge bezüglich der Verbesserung der Evaluationsqualität führen. Zudem fällt auf, dass Auftraggeber häufig in Diskussionen mit Evaluierenden verwickelt scheinen und dabei u.a. Verbesserungsmöglichkeiten besprechen, was die häufigen Nennungen verdeutlichen (neutrale Diskussion von Schlussfolgerungen, Diskussion der Ergebnispräsentation zur Verbesserung des Zielgruppenverständnis sowie der Evaluationsmethoden zur Verbesserung der Evaluationsqualität). Während 76 Prozent der Auftraggeber mehr als einmal den Evaluierenden gezeigt haben, welche Punkte verbessert werden können (*Betterment)*, haben sich 90 Prozent generell mehr als einmal für die Optimierung der Evaluationsqualität eingesetzt (*Support)*. Die Beeinflussungsformen *Betterment* und *Support* sind bei den US-Auftraggebern vergleichsweise stärker vertreten als *Distortion* und *Undermining*, was auf den Effekt der sozialen Erwünschtheit zurückgeführt werden kann. Dennoch fällt auf, dass die indirekte, implizite Beeinflussungsform sowohl bei der konstruktiven als auch der destruktiven Beeinflussungsart überwiegt. Die Frage nach möglichen Gründen für die dominierende indirekte Einflussnahme der Auftraggeber bleibt offen und kann durch zukünftige Forschung untersucht werden. Anhand eines Ländervergleichs könnten bspw. länderspezifische (kulturelle) Eigenheiten als De-

terminanten identifiziert werden. Die vorliegende Untersuchung der konstruktiven Beeinflussungsart liefert somit erstmalige Erkenntnisse über die konstruktive Beeinflussungsart und verdeutlicht die Wichtigkeit der Integration dieser Dimension für zukünftige Forschung. Abgeleitet von den Befunden stellen sich die Fragen: wie sind die konstruktive und destruktive Einflussnahme relativ zueinander bei den Auftraggebern ausgeprägt? Gibt es dabei innerhalb des Evaluationsprozesses situationsspezifische Unterschiede?

In Bezug auf die Berufserfahrung zeigt sich, dass die Mehrheit der US-Auftraggeber eher unerfahren sind, wobei zwar zwei Drittel der Auftraggeber bis zu 20 Evaluationen durchgeführt haben, aber 57 Prozent der Auftraggeber nur bis zu fünf Jahre Erfahrung in der Evaluationsvergabe aufweisen. Die gerechnete Zusammenhangsmasse zur Überprüfung der H7 liefern Hinweise darauf, dass die Berufserfahrung positiv – statt wie angenommen negativ – mit der konstruktiven Beeinflussungsart zusammenhängt. So gibt es Hinweise darauf, dass je mehr Berufserfahrung die Auftraggeber in ihrer Tätigkeit aufweisen, desto stärker ihre konstruktive Beeinflussungsart ausgeprägt ist. Trotz signifikanter Werte wurde H7 folglich abgelehnt. Somit kann für den Evaluationskontext nicht bestätigt werden, dass Auftraggeber aufgrund ihres Informationsdefizits (wenig Berufserfahrung) durch Investitionen in Form konstruktiver Einflussnahme dem Risiko entgegenwirken, dass Evaluierende entgegen ihren Interessen agieren. Dass H7 nicht zutrifft, scheint aber aufgrund verschiedener Gründe legitim. Der Befund deutet auf einen möglichen Spezialfall der PAT hin, der gerade darin liegt, dass nicht die Evaluierenden (*Agents*), sondern die Auftraggeber (*Principals*) über mehr Informationen verfügen und die Informationsasymmetrie gespiegelt wird. Damit zeigt sich, dass das Abhängigkeitsverhältnis zwischen Auftraggebern und Evaluierenden nicht nur unidirektional von einer Dependenz geprägt ist, sondern wechselseitig stattfindet und damit durch eine gewisse Interdependenz beider Akteure charakterisiert ist (Widmer, 2012, S. 131). Dies widerspricht der PAT hinsichtlich des Aspekts, dass sie vordergründig von einem Informationsvorteil seitens der Evaluierenden bezüglich der sachlichen Aufgabenbearbeitung ausgeht (Oehlrich, 2016, S. 121), wobei sich dieser Spezialfall auf den Informationsvorsprung der Auftraggeber in Bezug auf den Evaluationsgegenstand und -kontext bezieht und somit mit der sachlichen Aufgabenbearbeitung verknüpft ist. Dieser Informationsvorteil der Auftraggeber kann sich auch in den Evaluations- oder Methodenkompetenzen äussern. Die Aussage eines Befragten, dass der „Missbrauch oder die Fehldarstellung von Befunden ausserhalb des

Kontextes sowie die breite Verallgemeinerung auf Basis von kleinen und verzerrten Kohorten [durch den Evaluierenden]“ zu einem Konflikt führte, deutet auf eine gewisse Inkompetenz seitens des angesprochenen Evaluierenden hin. Gerade vor dem Hintergrund, dass Evaluierende hierzu nicht alles wissen (Perrin, 2018, S. 2), macht der signifikante positive Zusammenhang der H7 Sinn, dass dieser Informationsmangel durch die Unterstützung der Auftraggeber behoben werden kann. Die Berufserfahrung ermöglicht den Auftraggebern diese unterstützende Hilfe in Form einer konstruktiven Beeinflussungsart anzubieten. Die ergänzende Hilfe der Auftraggeber wird auch durch die Aussage unterstrichen, bei welcher eine Person auf die Schwierigkeit hinwies „Evaluierende zu finden, die sowohl über kontextuelle Expertise in die Perspektive der Programmbegünstigen als auch über methodische Expertise verfügen“. Das Mehrwissen seitens der Auftraggeber wurde auch bei Stockmann et al. (2011, S. 55) diskutiert, ohne aber die damit zusammenhängende Einflussnahme bezüglich ihrer Zulässigkeit zu bewerten. Neben der Berufserfahrung existiert zudem ein signifikanter, positiver Zusammenhang zwischen der Vertrautheit mit Evaluationsstandards und der konstruktiven Beeinflussungsart, wobei H6 vorläufig bestätigt wurde. Auch hier spiegelt sich die Vertrautheit in einer relativ kleineren Informationsasymmetrie. Sowohl die Berufserfahrung als auch die Vertrautheit mit Evaluationsstandards kommen nicht nur insgesamt dem Evaluationsprozess in Form einer konstruktiven Beeinflussung und nicht zuletzt der damit verbundenen Unabhängigkeit der Evaluation als Grundlage für EBP zugute, sondern nützen auch den Auftraggebern als *Principals* selbst. Durch die konstruktive Beeinflussungsart investieren die Auftraggeber nicht nur in das Beziehungsverhältnis mit den Evaluierenden, sondern tragen bspw. durch die frühzeitige Lieferung von relevanten Informationen dazu bei, dass Evaluierende nicht nur ihre eigenen Ziele verfolgen, sondern bevorzugt diejenigen die im Interesse der Auftraggeber – und hoffentlich auch der Evaluation – stehen. Dass die Evaluationsqualität ein Anliegen aufseiten der Auftraggeber ist, verdeutlichen verschiedene Befunde der Studie. 81 Prozent der Auftraggeber haben die ungenügende Evaluationsqualität als Unzufriedenheitsgrund angegeben, was sich wie erwähnt im Ländervergleich für die Schweiz bestätigt. Zudem haben 96 Prozent der US-Auftraggeber mehr als einmal einen konstruktiven Dialog zur Verbesserung der Evaluationsqualität geführt. Auch mindestens mehr als einmal haben 88 Prozent ihre Ideen mit Evaluierenden zur Qualitätsverbesserung ausgetauscht und 82 Prozent die Evaluationsmethoden zur Verbesserung der Evaluationsqualität diskutiert. Generell kann die Forschungsfrage *welche Rolle die Berufserfahrung und die Vertrautheit mit Evaluationsstandards in der Einflussnahme der*

Auftraggeber auf den Evaluationsprozess spielen, somit sicher beantwortet werden. Sowohl die Berufserfahrung als auch die Vertrautheit mit Evaluationsstandards hängen positiv mit der konstruktiven Beeinflussungsart der Auftraggeber auf den Evaluationsprozess zusammen. Die statistisch signifikanten Zusammenhänge weisen darauf hin, dass je mehr Auftraggeber mit den Evaluationsstandards vertraut sind und je mehr Berufserfahrung sie in ihrer Tätigkeit aufweisen, desto stärker ist ihre konstruktive Beeinflussungsart ausgeprägt.

Generell stellt sich die Frage, ob die Interessen der Auftraggeber und Evaluierenden tatsächlich so unterschiedlich sind und sich der im Rahmen der PAT angenommene Interessenkonflikt in der Evaluationsbeziehung doch weniger ausgeprägt ist als vermutet. Darauf weisen v.a. die Befunde des relativ schwach ausgeprägten Konfliktverhältnisses hin. Aus einer evaluationsethischen Perspektive sollten beide Akteure die gleichen, höhergestellten Interessen der wissenschaftlichen Unabhängigkeit und Integrität sowie der Einhaltung von Evaluationsstandards und einer hohen Evaluationsqualität verfolgen. Möglicherweise äussert sich der Konflikt im Beziehungsverhältnis weniger aufgrund divergierender Interessen, sondern vielmehr aufgrund der Schwierigkeiten in der Zusammenarbeit. Obwohl 67 Prozent der Auftraggeber nie bis eher selten mit Schwierigkeiten konfrontiert sind, äusserst sich eine Hauptschwierigkeit der Auftraggeber der USA, aber auch der Schweiz darin, dass das Verständnis für die zu evaluierende Organisation sowie das gegenseitige Verständnis beider Akteure fehlt. Aus diesem fehlenden Verständnis können möglicherweise unterschiedliche Erwartungen erwachsen, die wiederum zur Unzufriedenheit führen. Für das Konfliktverhältnis wird abgeleitet angenommen, dass Auftraggeber zwar Schwierigkeiten in der Zusammenarbeit erleben, mit Evaluationen unzufrieden sind, sich jedoch weder mit Evaluierenden darüber austauschen noch die Probleme offen ansprechen. Die Befunde zeigen, dass keinem US-Auftraggeber jemals eine Druckausübung oder Beeinflussung unterstellt wurde, wobei sich dieser Trend auch für die Schweizer Auftraggeber bestätigt. Im Umgang mit Divergenzen kann im Ländervergleich für die US-Evaluationslandschaft eine relativ „zurückhaltendere Kultur“ als für die Schweiz beobachtet werden. Dies zeigt sich einerseits an der relativ höheren Kompromissrate von 70 Prozent in den USA, welche die Kompromissfindung bei Änderungsvorschlägen wiederspiegelt. Andererseits kann dies am relativ kleineren Widerstand seitens der Evaluierende beobachtet werden, welche Änderungen entgegen des Auftraggeberwillens nicht vornehmen. Diese Befunde überraschen vor dem Hintergrund der Befunde von Morris

und Clark (2012, S. 61), die zeigen, dass 42 Prozent der Evaluierende jemals unter Druck gesetzt wurden Evaluationsresultate falsch darzustellen. Die widersprüchlichen, optimistischen Angaben der US-Auftraggeber könnten einerseits mit Rationalisierungsprozessen und Effekten der sozialen Erwünschtheit zusammenhängen, wodurch die Auftraggeber die tatsächliche Situation positiver darstellten, um ihr ethisch korrektes Image zu schützen. Ähnliche Verhaltensweisen wurden bereits bei Morris und Clark (2012, S. 67) für die Evaluierenden angenommen. Andererseits kann sich der Befund der fehlenden Unterstellung auch durch die Kultur der US-Evaluierenden erklären lassen. Evaluierende scheinen als wenig optimistisch eingestellt was die Prävention von Druckausübungen betrifft (Pleger et al., 2016, S. 12). Daraus kann abgeleitet werden, dass Evaluierende der USA weniger dazu neigen, die Druckausübung anzusprechen und dabei einen Konflikt mit dem Auftraggeber zu riskieren, was wiederum Auswirkungen auf mögliche Nachfolgeaufträge haben könnte. Wie auch bei Morris und Clark (2012) erwähnt sowie Pleger und Hadorn (2018) diskutiert, wird angenommen, dass Evaluierende durch die finanziellen Abhängigkeit von Evaluationsaufträgen eher bereit sind sich die Druckausübungen gefallen zu lassen oder Kompromisse einzugehen, um das Risiko eines zukünftigen Auftragsverlusts zu minimieren. Dennoch wird klar, dass es analog zur Schweiz auch der US-Evaluationslandschaft einer effektiven Konfliktkommunikation mangelt (Ebd., 2018, S. 14). Unabhängig davon kann der Befund, dass fast keine Änderungen vorgenommen wurden, aus ethischer Sicht als positiv bewertet werden. Ebenso deutet der Befund darauf hin, dass Evaluationsstandards zur Qualitätssicherung beitragen und Evaluierende diese befolgen, was sich mit den Befunden von Pleger et al. (2016) deckt.

Wie das Beziehungsverhältnis zwischen Auftraggebern und Evaluierenden ausgeprägt ist, ist auch vor dem Hintergrund relevant, dass divergierende Interessen automatisch vom Idealzustand der Evaluationen abweichen müssten. Diese Abweichung würde dann in der Verletzung der Unabhängigkeit seitens der Auftraggeber oder eben auch der Evaluierenden selbst bestehen. Die Verletzung der Unabhängigkeit seitens der Aufraggeber würde sich aufgrund der vorliegend untersuchten destruktiven Beeinflussungsart oder eines als destruktiv angenommenen Anreizsystems verstehen. Wie die vorgeschlagenen präventiven Massnahmen unterstreichen, soll ein gegenseitiges Verständnis der Akteure damit

gefördert werden, dass wie die der Arbeit zugrundeliegende Definition der Unabhängigkeit[12] (siehe Kapitel 2.2) vorschlägt, Interessenkonflikte in einer offenen und ehrlichen Art angegangen werden. Damit Evaluierende frei und unabhängig arbeiten können, werden destruktive Beeinflussungsformen abgelehnt und solche konstruktiver Art begrüsst. Das damit verbundene Ziel liegt darin, die Einflussformen negativer, destruktiver Art abzuschwächen und die positiver, konstruktiver Art zu verstärken (Pleger & Sager, 2018, S. 172). Die Befunde zeigen, dass alle US-Auftraggeber schon mehr als einmal frühzeitig alle für die Evaluation relevanten Informationen geliefert haben, um die Arbeit des Evaluierenden und somit die Unabhängigkeit von Evaluationen zu unterstützen. Trotzdem wird die allgemeine Verletzung der Unabhängigkeit von Evaluationen aufgrund der Befunde zur destruktiven, negativen Einflussnahme des Auftraggebers auf den Evaluationsprozess und der fehlenden, offenen Kommunikationskultur für die US-Evaluationslandschaft abgeleitet. Wie bereits erwähnt, stellt sich dabei jedoch die Frage inwiefern die destruktive in Relation zur konstruktiven Einflussnahme steht und ob letztere die Wirkungen der destruktiven Einflussnahme ggf. abschwächen oder gar eliminieren kann. Folglich lässt sich die Unabhängigkeit von Evaluationen in den USA nur aufgrund von Annahmen einschätzen und somit nicht sicher bewerten. Die Forschungsfrage *wie Auftraggeber die Wichtigkeit von unabhängigen Evaluationen beurteilen und die Unabhängigkeit von Evaluationen, sowie ihre eigene Einflussstärke wahrnehmen,* lässt sich anhand der deskriptiven Statistiken ableiten. Neben den diskutierten Befunden, dass den Auftraggebern die Evaluationsqualität ein Anliegen ist, zeigt sich, dass 55 Prozent der Auftraggeber die Unabhängigkeit von Evaluationen, d.h. dass Evaluationen generell ohne Einfluss aufseiten der Auftraggeber durchgeführt werden, als sehr wichtig einordnen. Davon halten dennoch 13 Prozent die Unabhängigkeit von Evaluationen als tendenziell unwichtig und 87 Prozent als tendenziell wichtig. Was die Beweggründe der Auftraggeber sind, die Unabhängigkeit als eher unwichtig einzuschätzen, bleibt offen. Diese Unwichtigkeit lässt sich ggf. mit einem differenzierten Evaluationsverständnis erklären, was durch die ohnehin tiefe Bekanntheit der nationalen „Program Evaluation Standards“ von 70 Prozent der Befragten durchaus möglich scheint. Die Mehrheit der Auftraggeber neh-

[12] Der Standard zur Unabhängigkeit von Evaluierenden wird wie folgt definiert (OECD, 2010, S. 11): “Evaluators are independent from the development intervention, including its policy, operations and management functions, as well as intended beneficiaries. Possible conflicts of interest are addressed openly and honestly. The evaluation team is able to work freely and without interference. It is assured of co-operation and access to all relevant information.”

men die Evaluationen generell als sehr unabhängig wahr (Werte 4-5 von 5). Die Einflussstärke der Auftraggeber auf den Evaluationsprozess wird von der Mehrheit der Befragten als tief bis mittelmässig eingeordnet. Dennoch haben 17 Prozent ihren Einfluss mit dem hohen Wert von 8 (von insgesamt 10) eingestuft. Dabei muss beachtet werden, dass nicht ein negativer Einfluss per se gemessen wurde, sondern die Einflussstärke, je nach Auffassung der Befragten, auch eine positive Einflussnahme miteinschliesst. Generell sind diese Befunde mit Vorsicht zu interpretieren, da Effekte der sozialen Erwünschtheit sehr wahrscheinlich sind und es sich bei den Daten lediglich um Selbsteinschätzungen der Befragten handelt.

Die Forschungsfrage *was von den Auftraggebern wahrgenommene Gründe für das konfliktgeprägte Verhältnis sind und welche präventiven Massnahmen vorgeschlagen werden, um ein fruchtbares Umfeld für aussagekräftige Evaluationen zu schaffen,* wird folgend beantwortet. Besonders häufig wurde das Fehlen von wichtigen Kompetenzen, die mangelhafte Qualität der Evaluationsresultate und das fehlende Verständnis der Anforderungen als Konfliktgründe genannt. Auch wurden interpersonale Kompetenzen und mangelnde kommunikative und proaktive Verhaltensweisen genannt. Dabei überrascht die meistgenannte (82%) präventive Massnahme nicht, dass ein grösseres gegenseitiges Verständnis der Zielsetzung, Zweckbestimmung und Funktionen geschaffen werden sollte. Danach wurde die engere Zusammenarbeit beider Parteien (55%) genannt. Dabei wird deutlich, dass die Auftraggeber die erwähnte mangelhafte Kommunikation zwischen den Evaluierenden erkannt haben und für ein grösseres gegenseitiges Verständnis plädieren. Das fehlende (gegenseitige) Verständnis wurde gleichzeitig als Hauptschwierigkeit in der Zusammenarbeit mit Evaluierenden identifiziert, weshalb dieser Vorschlag nicht überrascht. Wird dahingehend unterschieden welche Massnahmen von Auftraggebern mit resp. ohne Konflikt genannt wurden, zeigt sich, dass Personen ohne Konflikt die jeweiligen Massnahmen tendenziell häufiger ausgewählt haben als diejenigen mit Konflikt. Die neutrale Intervention von Dritten sowie die meistgenannte Prävention der Schaffung eines grösseren gegenseitigen Verständnisses wurden häufiger von Personen mit Konflikt ausgewählt. Diese Befunde deuten darauf hin, dass Auftraggeber ohne Konflikt möglicherweise relativ positivere Erfahrungen mit Evaluierenden gemacht haben und dabei gerade die Massnahmen angegeben haben, die sie bereits anwenden. Für diejenigen mit Konflikt wird angenommen, dass sie bereits etwas resigniert haben und die Problematik weniger bei sich selbst sehen, was den Attributionsprozessen zugeschrieben werden kann

(Kelley, 1973). Entsprechend weisen sie auf die Intervention von Dritten hin oder betonen den Einsatz beider Akteure zum gegenseitigen Verständnis. Im Kontext der PAT birgt diese Intervention von Dritten eine Art Ausweg aus dem Dilemma der im Beziehungsverhältnis inhärenten Macht- und Interessenasymmetrien. Durch die unabhängige Mediatorenstelle können diese Strukturen aufgebrochen und in ein Gleichgewicht gebracht werden, während optimalerweise konstruktive Einflüsse im Evaluationsprozess gefördert und destruktive Einflüsse minimiert werden (Pleger & Hadorn, 2018, S. 15). Der Befund überrascht insofern, dass diese Massnahme unter den Schweizer Auftraggebern sehr unpopulär war und dafür vorliegend keine Erklärung gefunden wird. Jedoch kann abgeleitet angenommen werden, dass die Konflikterfahrung als Lernprozess fungieren kann, der zur Erkenntnis der Auftraggeber führt, dass die Involvierung von Dritten zielführend sein kann.

Zusammenfassend zeigen die Befunde, dass die Bandbreite möglicher Beeinflussungsformen – ob konstruktiv oder destruktiv – sehr gross ist. Generell wurde ein positiver, statistisch signifikanter Zusammenhang sowohl zwischen der Unzufriedenheit und der destruktiven Einflussnahme als auch dem Konfliktverhältnis und dem Anreizsystem identifiziert. Für die konstruktive Einflussnahme auf den Evaluationsprozess wurden positive signifikante Zusammenhänge mit der Vertrautheit als auch der Berufserfahrung gefunden. Analog zur Schweizer Studie überrascht für den US-Evaluationskontext, dass zwar generell ein konfliktgeprägtes Verhältnis vorherrscht, jedoch keinem Auftraggeber jemals eine Druckausübung unterstellt wurde. Zur zukünftigen Konfliktprävention- und reduktion wird insb. auf eine ehrliche und offene Konfliktkommunikation plädiert, die auch das gegenseitige Verständnis zwischen den Auftraggebern und Evaluierenden fördert und damit Unzufriedenheiten und Schwierigkeiten in der Zusammenarbeit reduziert. Zudem wurde ein alarmierendes Informationsdefizit aufseiten der Auftraggeber hinsichtlich der Bekanntheit der nationalen „Program Evaluation Standards“ identifiziert. Nach dieser ausführlichen Diskussion werden im nächsten Kapitel schlussfolgernd die Implikationen und Limitationen erörtert und den Blick auf zukünftige Forschung gerichtet.

6 Schlussfolgerungen

In diesem Kapitel werden die Schlussfolgerungen anhand der Implikationen und Limitationen erörtert, wobei die Masterarbeit mit einem Ausblick abgerundet wird.

6.1 Implikationen

Die vorliegende explorative Studie gilt als erster Versuch einen umfassenden Überblick über die Auftraggeberperspektive des Evaluationskontextes der USA zu schaffen. Dabei verfolgte die Studie nicht das Ziel der Repräsentativität, sondern vielmehr einen deskriptiven Einblick darüber zu geben, wie die Unabhängigkeit von Evaluationen aus Auftraggebersicht in den USA beurteilt wird. Genauer wurde untersucht, welche Rolle die Eigenschaften der Auftraggeber und ihre Beziehung zu Evaluierende in der Beeinflussung des Evaluationsprozesses spielen. Darüber hinaus bietet die Studie wertvolle Erkenntnisse darüber wie die Auftraggeber im digitalen Zeitalter von Problematiken wie Spam und erhöhten Sicherheitsbestimmungen zu erreichen sind und präsentiert Lösungsvorschläge wie diese Herausforderungen während zukünftigen Datenerhebungen gemeistert werden können. Damit fungiert die Studie als idealer methodischer Anknüpfungspunkt für zukünftige Forschung in diesem Bereich. Die Befunde liefern zudem einen einmaligen sowie facettenreichen Einblick in den Erfahrungsschatz, die Wahrnehmungen und Einstellungen von Auftraggebern in der USA mit besonderem Fokus auf deren Eigenschaften und dem Beziehungsverhältnis zu Evaluierenden. Durch den Vergleich zur Schweizer Studie von Pleger und Hadorn (2018) fungiert die Studie zur Validitätsprüfung jener Befunde. Auch wurde die Studie sowohl methodisch durch eine höhere Standardisierung der Datenerhebung und der Inklusion des BUSD-Modells zur Operationalisierung der Konstrukte als auch theoretisch durch einen stärkeren Fokus auf der PAT erweitert.

Die Befunde zeigen, dass die Bandbreite möglicher Beeinflussungsarten – ob konstruktiv oder destruktiv – sehr gross ist. Dabei ist es wichtig zwischen beiden Beeinflussungsformen sowie der indirekten und direkten Beeinflussung zu unterscheiden. Diese Unterscheidung ist notwendig damit definiert werden kann, ob Änderungen zur Verbesserung oder Verzerrung von Evaluationsresultaten führen und dadurch die Unabhängigkeit von Evaluationen aufrechterhalten oder gefährden. Das BUSD-Modell liefert dabei die optimale Basis zur Weiterentwicklung zukünftiger Forschung, aber auch zur praktischen Anwendung sowohl für Evaluierende als auch, wie durch die Studie erweitert, für Auftraggeber. Letztere werden in ihrer Wahrnehmung unterstützt, ob der ausgeübte Einfluss konstruktiver oder destruktiver Natur ist. Diese Unterscheidung hat auch gesamtgesellschaftliche Implikationen und scheint unerheblich für das demokratische Outcome von Evaluationen, da daraus Handlungsempfehlungen zur Prävention von negativem Einfluss resp. zur Förderung von positivem Einfluss abgeleitet werden können (Pleger & Sager, 2018,

S. 168). Indem mögliche Ursachen für das während dem Evaluationsprozess inhärente Konfliktpotenzial zwischen dem Auftraggeber und Evaluierenden sowie die Sichtweisen der Auftraggeber erforscht wurden, können für die Praxis mögliche Handlungsempfehlungen und präventive Massnahmen abgeleitet werden. Diese fungieren zur Bereicherung des gegenseitigen Austauschs zwischen Auftraggebern und Evaluierenden und optimalerweise zur Gewährleistung und Verbesserung der Unabhängigkeit von Evaluationen. Dabei sind die drei nachfolgenden Handlungsempfehlungen zentral. Die erste Handlungsempfehlung liegt in der Schaffung eins grösseren, gegenseitigen Verständnisses zwischen Auftraggebern und Evaluierenden. Dabei kann eine gründliche Bedarfs- und Erwartungsabklärung zu Beginn einer Evaluation mögliche Missverständnisse beseitigen und eine Grundlage für die zukünftige Zusammenarbeit schaffen. Bei einem besonders konfliktgeprägten Verhältnis empfiehlt sich zudem die neutrale Intervention von Dritten in Form einer unabhängigen Mediatorenstelle. Zweitens wird zur Prävention und Reduktion von Konflikten eine offene Kommunikationskultur zwischen beiden Akteuren empfohlen, die nicht nur bei Konflikten, sondern schon vorher Unstimmigkeiten, Unzufriedenheit und Schwierigkeiten offen im konstruktiven Dialog anspricht und somit das gegenseitige Verständnis im Dialog fördert. Drittens soll das alarmierende Informationsdefizit der Auftraggeber bezüglich der nationalen „Program Evaluation Standards" reduziert werden. Obwohl die Pflicht auch bei den Auftraggebern selbst liegt, sich über übliche Richtlinien und Standards genügend zu informieren, richtet sich dieser Aufruf ebenso an die AEA als zentrale Institution, deren Mission u.a. in der Verbesserung der Evaluationspraxis liegt. Während die AEA (2019a) grossen Wert auf die Inklusion und Diversität ihrer Mitgliedschaft legt, sollten vermehrt auch die Auftraggeber im Austausch miteinbezogen werden, sodass die Relevanz und Dringlichkeit der Evaluationsstandards den Auftraggebern vermittelt werden kann. Dies könnte bspw. in Form breitangelegter Informationskampagnen und Tagungen stattfinden aber auch in der individuellen Leistung von Evaluierenden, die Auftraggeber diesbezüglich zu schulen, was längerfristig zur kontinuierlichen Reduktion des Informationsdefizits der US-Auftraggeber beitragen könnte. Das Verständnis der Auftraggeberseite ermöglicht somit eine vollständigere Beschreibung der wechselseitigen Evaluationsbeziehung im komplexen Kontext des Evaluationsprozesses. Der Grad der Unabhängigkeit von Evaluationen kann schliesslich nur dann verbessert werden, wenn verstanden wird, ob und inwiefern das Bewusstsein bezüglich unethischer Verhaltensweisen aufseiten beider Perspektiven ausgeprägt ist (Pleger & Hadorn, 2018, S. 4). Picciotto (2019, S. 95) folgert zudem, dass die Unabhängigkeit von Evaluationen

mit der Professionalisierung der Evaluationsdisziplin einhergeht. Die Integrität von Evaluationsprozessen kann gewahrt werden, wenn die notwendige berufliche Autonomie mittels Akkreditierungssystemen und ethisch einwandfreien, überarbeiteten Evaluationsrichtlinien erlangt wird. Damit kann auch das Verfolgen von Eigeninteressen vermieden werden, um nicht zuletzt die gesellschaftliche Wirkung der Evaluation zu erhöhen. Die für die Studie identifizierten Limitationen werden im nächsten Abschnitt erläutert.

6.2 Limitationen

Im Rahmen der Datenerhebung kam es zu Schwierigkeiten bei der Rekrutierung von Befragungsteilnehmenden, was vordergründig auf den tiefen Organisationsgrad der Auftraggeber von Evaluationen im Allgemeinen zurückgeführt werden kann. Da keine öffentlich zugänglichen Verzeichnisse von Auftraggebern der gesamten Evaluationslandschaft oder einzelner Sektoren vorliegen, sind sie schwierig erreichbar. Andererseits deuten die Erfahrungen während der Datenerhebung darauf hin, dass eine Hauptproblematik darin liegt, dass viele E-Mails aufgrund von Spamfiltern und internen Sicherheitsmassnahmen erst gar nicht bei den angeschriebenen Personen ankommen. Dazu wird vermutet, dass die ZHAW als durchführende Institution unter den US-Auftraggebern nicht bekannt ist und zur Rekrutierung idealerweise ihnen vertraute Organisationen oder Personen nötig sind, um den Umfragelink zu verbreiten oder potenzielle Respondenten zur Studienteilnahme zu motivieren. Zudem ermöglicht der von Qualtrics generierte anonymisierte Link weder systematischen Reminder-E-Mails noch Rückschlüsse auf die Rücklaufquote und Grundgesamtheit, was eine weitere Limitation darstellt. Trotz vielfältiger Rekrutierungsmassnahmen resultierte somit eine geringe Samplegrösse. Dennoch zeugt die frühzeitige Abbruchrate bei der Online-Befragung davon, dass die angeschriebenen Personen generell interessiert sind und ihre Aufmerksamkeit erweckt wurde, jedoch nicht genügend motiviert zur Weiterbearbeitung und Beendigung der Befragung waren. Folglich könnte ein Selection Bias vorliegen, dass das untersuchte Sample v.a. aus Auftraggebern besteht, die besonders am Forschungsthema interessiert oder generell stärker engagiert sind als andere Auftraggeber. Obwohl die Anzahl tatsächlich auswertbarer Fälle im Verhältnis zu den aufgewendeten Zeitressourcen für die Rekrutierung relativ tief ausfällt, gibt es Hinweise auf die Validität der Relevanz der Respondenten. Dies resultiert daraus, dass die Rückmeldungen gewisser Persönlichkeiten, die für die Teilnahme der Online-Befragung geeignetste Person intern zu identifizieren, von einer gewissen Relevanzabklärung zeugen. Für den Ländervergleich stimmen zudem die Fallzahlen oft mit der Schweizer Studie überein, was eine geeignete Basis für den Vergleich darstellt.

Abschliessend gilt zu berücksichtigen, dass aufgrund der sensiblen Thematik der Unabhängigkeit von Evaluationen Effekte der sozialen Erwünschtheit bei den Befragten aufgetreten sind, welche die Resultate systematisch verzerren (Pleger et al., 2016, S. 4). Die Zusammenhangsberechnungen konnten nicht für diese Effekte kontrolliert werden, weshalb die Resultate dahingehend mit Vorsicht zu interpretieren sind. Die Variable der *sozialen Erwünschtheit* bestätigt, dass 75 Prozent der Befragten eher zu sozial erwünschten Antworten neigen, während dies für 25 Prozent eher nicht zutrifft. Dabei liegen keine Angaben vor, dass die soziale Erwünschtheit für die Befragten gar nicht ausgeprägt ist.
Zudem ist die Durchführung einer repräsentativen Studie zum heutigen Zeitpunkt aufgrund der fehlenden Grundgesamtheit der US-Auftraggeber unmöglich, was zwar eine Limitation darstellt, jedoch nicht das Ziel dieser explorativen Forschung war (vergleiche auch Stockmann et al., 2011, S. 46). Somit können die Befunde weder generalisiert noch können inferenzstatistische Aussagen getroffen werden. Weiter beziehen sich die Befunde nicht auf Einzelfälle, sondern auf die Gesamtheit der befragten US-Auftraggeber.
Eine weitere Limitation liegt in der indirekten Messung der Informationsasymmetrie, welche vorliegend durch die Variablen der *Berufserfahrung* und *Vertrautheit* operationalisiert wurde. Um die Informationsasymmetrie direkt zu messen, müsste der tatsächliche Informationsstand hinsichtlich des ganzheitlichen Auftragsverhältnis mitsamt Evaluationsgegenstand und -kontext sowohl der Auftraggeber als auch der Evaluierende bekannt sein, um dann die Differenz zu identifizieren, die dann wiederum Aussagen über die jeweilige Informationsasymmetrie zulässt. Die Annahme der PAT, dass die Informationsasymmetrie zulasten der *Principals* ausfällt, impliziert eine weitere Limitation, da wie diskutiert, die Auftraggeber bezüglich Evaluationsgegenstand und -kontext über einen Informationsvorteil verfügen können. Die Informationsasymmetrie kann somit situationsabhängig sowohl für den *Principal* als auch den *Agent* variieren, was im Rahmen zukünftiger Forschung vertieft untersucht werden kann. Weitere Forschung müsste sich auch verstärkt mit der Untersuchung der jeweiligen (divergierenden) Interessen beider Akteure auseinandersetzen, die den Interessenkonflikt des Principal-Agent-Verhältnisses prägen. Die vorliegend diskutierten Interessen berufen mehrheitlich auf Annahmen resp. wurden zusammengefasst als Interessenkonflikt gemessen. Eine Limitation des BUSD-Modells liegt darin, dass es als konzeptionelles Modell keine Erklärungen darüber bietet wie konstruktive und destruktive Beeinflussungsformen die Evaluationsqualität beeinflussen (Pleger & Sager, 2018, S. 172). Die erweiterte Untersuchung welche Wirkung die

Einflussnahme von Auftraggebern nicht nur auf den Evaluationsprozess, sondern auf die Evaluationsqualität und die Unabhängigkeit von Evaluationen hat, zeugt vor dem Hintergrund des EBP von demokratiepolitischer und somit gesamtgesellschaftlicher Relevanz. Der letzte Abschnitt rundet die Studie mit einem Ausblick für zukünftige Forschung ab.

6.3 Ausblick

Analog zur ländervergleichenden Studie von Pleger et al. (2016) mit Fokus auf der Druckausübung auf Evaluierende könnte zukünftige Forschung zu einem vollständigen Portrait der internationalen Evaluationslandschaft beitragen, indem Studien zur Auftraggeberseite – ergänzend zur vorliegenden Studie und derjenigen von Pleger und Hadorn (2018)– in Grossbritannien und Deutschland durchgeführt werden. Diese Studien würden dann als ideale Basis für eine ländervergleichende Studie fungieren, welche die Befunde der Evalueriенden- und Auftraggeberperspektive zusammenführt. Gerade durch die integrative Betrachtung beider Seiten können Gemeinsamkeiten und Divergenzen identifiziert werden, um daraus sowohl theoretische Rückschlüsse als auch Handlungsempfehlungen für die Praxis abzuleiten. Dies deckt sich mit der Aussage Picciotto's (2019, S. 95), dass das Evaluationsgebiet stärker internationalisiert werden muss, um den weltweiten Fortschritt zu unterstützen. Zudem sollten Evaluationen in alle gesellschaftlichen Sektoren expandieren, da die nationalen aber auch sektorspezifischen Grenzen zwischen den Evaluationen im privaten, öffentlichen und gemeinnützigen Bereich zunehmend aufbrechen. Um mit diesen Tendenzen schrittzuhalten, scheint es unerheblich, dass sich die Forschung nicht nur über die Ländergrenzen hinwegbewegt, sondern auch wie die vorliegende Studie sektorunabhängige Evaluationsbeziehungen untersucht. Zudem wurden bisher v.a. Zusammenhänge untersucht, wobei sich zukünftige Forschung vermehrt auf die Untersuchung von Wirkungsbeziehungen annehmen und die vorliegenden Hypothesen diesbezüglich erweitern und analysieren kann. Bislang wurde v.a. auf das Beziehungsverhältnis zwischen *Principal* und *Agent* fokussiert, wobei dieses Verhältnis auch von anderen Faktoren wie weiteren Stakeholdern geprägt ist, die direkt im Evaluationsprozess involviert sind. Somit könnte sich zukünftige Forschung der komplexeren Umwelt von mehreren *Principals* annehmen, die noch besser der Realität entspricht (Balago, 2014, S. 251). Weiter könnte das Zusammenspiel zwischen *Principal* und *Agent* anhand des „Five Forces Framework" für den US-Evaluationsmarkt untersucht werden, wobei Evaluationen als Industrie mit einer Marktplatzdynamik verstanden und unter diesem Gesichtspunkt näher beleuchtet werden könnten (Furubo & Stame, 2019, S. 246–247). Eine zentrale Erweiterung würde die bereits erwähnte Untersuchung der Wirkung von Beeinflussungsformen

auf die Evaluationsqualität darstellen (siehe Kapitel 6.2). In Anbetracht der schwierigen Messbarkeit der Unabhängigkeit von Evaluationen, die meist auf Selbsteinschätzungen der Befragten basieren und mit Effekten der sozialen Erwünschtheit behaftet sind, würde sich dafür die Konzeption eines Index anbieten. Ein solcher Index würde nicht nur theoretisch verwendet, sondern v.a. in der Praxis eingesetzt werden können. Ein grosser Vorteil würde in der Vergleichbarkeit von Evaluationen hinsichtlich ihrer Qualität sowie Unabhängigkeit liegen. Durch dessen Verwendung könnte die Dringlichkeit und Aufmerksamkeit von unabhängigen Evaluationen in der Evaluationspraxis verstärkt und ein Orientierungsrahmen für Evaluierende und Auftraggeber geschaffen werden. Bei der Beurteilung der Unabhängigkeit von Evaluationen ist ergänzend festzuhalten, dass dafür der beabsichtigte Evaluationsnutzen mitberücksichtigt werden sollte. Während summative Evaluationen auf einen Leistungsnachweis abzielen und sich an externe Adressaten wie die Öffentlichkeit oder Politik richten, sind formative Evaluationen stärker lernorientiert und auf eine Verbesserung des Evaluationsgegenstandes ausgerichtet. Letztere richten sich somit primär an interne Adressaten wie Nutzende, die sich direkt mit dem Gegenstand befassen. Abhängig vom Adressatenkreis unterscheiden sich die Glaubwürdigkeitserfordernisse einer Evaluation, die in Zusammenhang mit der Unabhängigkeit von Evaluationen stehen. Bei internen Adressaten steht vielmehr die Vertrautheit mit dem Evaluationsgegenstand im Vordergrund, wobei bei externen Adressaten die Unabhängigkeit stärker gewertet wird (Widmer, 2012, S. 134–135). Zukünftige Forschung könnte sich auch dem von Picciotto (2019, S. 88) vorgeschlagenen Modell der unabhängigen demokratischen Evaluation annehmen. Das Modell ist v.a. dann erforderlich, wenn demokratische Evaluationsansätze behindert werden, indem sie von Auftraggebern streng kontrolliert werden oder Dissens nicht tolerieren. Evaluierende, welche nach diesem Modell agieren, würden Aufträge ablehnen, bei welchen direkt an die Entscheidungsträger berichtet werden muss, welche für die zu evaluierende Intervention verantwortlich sind. Stattdessen wird an eine übergeordnete Einheit mit einem gewissen Abstand zum Evaluationsgegenstand berichtet wie bspw. an eine NGO (Ebd., 2019, S. 94). Dabei könnte untersucht werden, welche Konsequenzen sich für das Beziehungsverhältnis zwischen Auftraggebern und Evaluierenden ergeben und ob sich die vorgeschlagene Neuausrichtung der Evaluationspolitik tatsächlich positiv auf die Unabhängigkeit von Evaluationen auswirkt.

Literaturverzeichnis

American Evaluation Association (2011). *Guiding principles*. Abgerufen am 23. Januar 2019, von https://www.eval.org/p/cm/ld/fid=51.

American Evaluation Association (2019a). About us. Abgerufen am 20. Januar 2019, von https://www.eval.org/p/cm/ld/fid=4.

American Evaluation Association (2019b). University programs. Abgerufen am 20. April 2019, von https://www.eval.org/p/cm/ld/fid=43.

Balago, G. S. (2014). A conceptual review of agency models of performance evaluation. *International Journal of Finance and Accounting*, *3*(4), 244–252.

Balzer, L. (2019). *Evaluation portal: Overview*. Abgerufen am 4. Oktober 2019, von https://www.evaluation.lars-balzer.name.

Barnett, C., & Camfield, L. (2016). Ethics in evaluation. *Journal of Development Effectiveness*, *8*(4), 528–534. https://doi.org/10.1080/19439342.2016.1244554

Benninghaus, H. (2007). *Deskriptive Statistik. Eine Einführung für Sozialwissenschaftler* (11. Auflage). Wiesbaden: VS Verlag für Sozialwissenschaften.

Brown, R. D., & Newman, D. L. (1992). Ethical principles and evaluation standards: Do they match? *Evaluation Review*, *16*(6), 650–663. https://doi.org/10.1177/0193841X9201600605

Bühl, A. (2014). *SPSS 22: Einführung in die moderne Datenanalyse* (14., aktualisierte Auflage ed., Vol. 4249, Pearson Studium - Scientific Tools). Hallbergmoos: Pearson.

Caers, R., Bois, C. D., Jegers, M., Gieter, S. D., Schepers, C., & Pepermans, R. (2006). Principal-agent relationships on the stewardship-agency axis. *Nonprofit Management and Leadership*, *17*(1), 25–47. https://doi.org/10.1002/nml.129

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. Auflage). New York: Psychology Press.

Council on Foundations (2019). *Listing of council member websites*. Abgerufen am 4. Januar 2019, von https://www.cof.org/members-directory/non-members.

Davis, J. H., Schoorman, F. D., & Donaldson, L. (1997). Toward a stewardship theory of management. *The Academy of Management Review*, *22*(1), 20–47. https://doi.org/10.2307/259223

Desautels, G., & Jacob, S. A. (2012). The ethical sensitivity of evaluators : A qualitative study using a vignette design. *Evaluation*, *18*(4), 437–450.

Diekmann, A. (2013). *Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen* (7. Auflage ed., Vol. 55678, Rowohlts Enzyklopädie). Reinbek: Rowohlt.

Dormann, C. F. (2017). *Parametrische Statistik: Verteilungen, maximum likelihood und GLM in R* (2. Auflage ed., Statistik und ihre Anwendungen). Berlin Heidelberg: Springer Berlin Heidelberg.

Eastmond, N. (1998). Commentary: When funders want to compromise your design. *American Journal of Evaluation*, *19*(3), 392–395.

Eisenhardt, K. M. (1989). Agency theory: An assessment and review. *The Academy of Management Review*, *14*(1), 57–74. https://doi.org/10.2307/258191

Forbes. (2019). The 100 largest U.S. charities. Abgerufen am 3. Juni 2019, von https://www.forbes.com/top-charities/list/#tab:rank.

Fortune. (2019). Fortune 500. Abgerufen am 3. Februar 2019, von http://fortune.com/fortune500/list.

Fox, C., Grimm, R., & Caldeira, R. (2017). *An introduction to evaluation*. Los Angeles: SAGE.

Furubo, J.-E., & Stame, N. (2019). *The evaluation enterprise. A critical view.* New York: Routledge.

Grossman, S. J., & Hart, O. D. (1983). An analysis of the principal-agent problem. *Econometrica*, *51*(1), 7–45. https://doi.org/10.2307/1912246

Head, B. W. (2008). Three lenses of evidence-based policy. *Australian Journal of Public Administration*, *67*(1), 1–11. https://doi.org/10.1111/j.1467-8500.2007.00564.x

Hemmerich, W. A. (2019). *Cronbachs Alpha: Auswerten und berichten.* Abgerufen am 6. April 2019, von https://statistikguru.de/spss/reliabilitaetsanalyse/auswerten-und-berichten-2.html.

Hochschule Luzern (2019). *Rangkorrelation*. Abgerufen am 25. Mai 2019, von https://www.empirical-methods.hslu.ch/entscheidbaum/zusammenhaenge/rang-korrelation/.

Holmström, B. (1979). Moral hazard and observability. *The Bell Journal of Economics*, *10*(1), 74–91. https://doi.org/10.2307/3003320

Jacob, S., & Boisvert, Y. (2010). To be or not to be a profession: Pros, cons and challenges for Evaluation. *Evaluation*, *16*(4), 349–369. https://doi.org/10.1177/1356389010380001

Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, *3*(4), 305–360.

Kaluza, B., Dullnig, H., & Malle, F. (2003). *Principal-Agent-Probleme in der Supply Chain: Problemanalyse und Diskussion von Lösungsvorschlägen* (Vol. Nr. 2003/03, Diskussionsbeiträge des Instituts für Wirtschaftswissenschaften der Universität Klagenfurt). Klagenfurt: Universität Klagenfurt, Institut für Wirtschaftswissenschaften.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, *28*(2), 107–128. https://psycnet.apa.org/doiLanding?doi=10.1037%2Fh0034225

Krüger, C., Borgmann, L., & Antonik, T. (2012). *Datenauswertung mit SPSS*. Abgerufen am 1. Mai 2019, von http://www.zhb.tu-dortmund.de/zhb/Row/Medienpool/Downloads/spss_1_2.pdf.

Mitnick, B. M. (1986). *The theory of agency and organizational analysis*. Gehalten am Annual Meeting der American Political Science Association, Washington, D.C.

Morris, M. (1999). Research on evaluation ethics: What have we learned and why is it important? *New Directions for Evaluation*, *1999*(82), 15–24. https://doi.org/10.1002/ev.1133

Morris, M. (2007). Foundation officers, evaluation, and ethical problems: A pilot investigation. *Evaluation and Program Planning*, *30*(4), 410–415. https://doi.org/10.1016/j.evalprogplan.2007.06.003

Morris, M. (2008). *Evaluation ethics for best practice: Cases and commentaries.* New York: Guilford Press.

Morris, M., & Clark, B. (2012). You want me to do what? Evaluators and the pressure to misrepresent findings. *American Journal of Evaluation*, *34*(1), 57–70. https://doi.org/10.1177/1098214012457237

Morris, M., & Cohn, R. (1993). Program evaluators and ethical challenges: A national survey. *Evaluation Review*, *17*(6), 621–642. https://doi.org/10.1177/0193841X9301700603

Morris, M., & Jacobs, L. R. (2000). You got a problem with that? Exploring evaluators' disagreements about ethics. *Evaluation Review*, *24*(4), 384–406.

Nutley, S., Morton, S., Jung, T., & Boaz, A. (2010). Evidence and policy in six European countries: diverse approaches and common challenges. *Evidence & Policy: A Journal of Research, Debate and Practice*, *6*(2), 131–144. https://doi.org/10.1332/174426410X502275

Organisation for Economic Co-Operation Development (2010). *Quality standards for development evaluation*. Paris: OECD Publishing.

Oehlrich, M. (2016). *Organisation. Organisationsgestaltung, Principal-Agent-Theorie und Wandel von Organisationen*. München: Verlag Franz Vahlen.

Perrin, B. (2018). How to manage pressure to change reports: Should evaluators be above criticism? *American Journal of Evaluation*. 1-22. https://doi.org/10.1177/1098214018792622

Picciotto, R. (2019). Is evaluation obsolete in a post-truth world? *Evaluation and Program Planning*, *73*, 88–96. https://doi.org/10.1016/j.evalprogplan.2018.12.006

Pleger, L. E., & Hadorn, S. (2018). The big bad wolf's view: The evaluation clients' perspectives on independence of evaluations. *Evaluation*, *24*(4), 1–19. https://doi.org/10.1177/1356389018796004

Pleger, L. E., & Sager, F. (2016a). Die Beeinflussung in der Evaluationstätigkeit in der Schweiz und was die SEVAL dagegen tun kann. *LeGes-Gesetzgebung & Evaluation*, *27*(1), 33–49.

Pleger, L. E., & Sager, F. (2016b). „Don't tell me cause it hurts" – Beeinflussung von Evaluierenden in der Schweiz. *Zeitschrift für Evaluation*, *15*(1), 23–59.

Pleger, L. E., & Sager, F. (2016c). *Evaluation and Independence. Existing evaluation policies and new approaches* [Bericht für UNDP Independent Evaluation Office

(IEO)]. Abgerufen am 25. November 2018, von http://web.undp.org/evaluation/evaluations/documents/Independence_of_Evaluation.pdf.

Pleger, L. E., & Sager, F. (2018). Betterment, undermining, support and distortion: A heuristic model for the analysis of pressure on evaluators. *Evaluation and Program Planning*, *69*(2018), 166–172. https://doi.org/10.1016/j.evalprogplan.2016.09.002

Pleger, L. E., Sager, F., Morris, M., Meyer, W., & Stockmann, R. (2016). Are some countries more prone to pressure evaluators than others? Comparing findings from the United States, United Kingdom, Germany, and Switzerland. *American Journal of Evaluation*, 1–14. https://doi.org/10.1177/1098214016662907

Pope, K. S., & Vetter, V. A. (1992). Ethical dilemmas encountered by members of the American Psychological Association: A national survey. *American Psychologist*, *47*(3), 374–411.

Posavac, E. J. (2014). *Program evaluation: methods and case studies* (8. Auflage). Harlow, Essex: Pearson.

Qualtrics (2019a). Avoid being marked as spam. Abgerufen am 4. Oktober 2019, von https://www.qualtrics.com/support/survey-platform/distributions-module/email-distribution/avoid-being-marked-as-spam/.

Qualtrics (2019b). Meet the software that powers more than 1 billion survey every year. Abgerufen am 5. Juli 2019, von https://www.qualtrics.com/uk/research-core/survey-software/?rid=ip&prevsite=de&newsite=uk&geo=CH&geomatch=uk.

Qualtrics (2019c). Question types. Abgerufen am 14. Mai 2019, von https://www.qualtrics.com/support/survey-platform/survey-module/editing-questions/question-types-guide/question-types-overview/.

Qualtrics (2019d). Social media distribution. Abgerufen am 15. April 2019, von https://www.qualtrics.com/support/survey-platform/distributions-module/social-media-distribution/.

Richter, R. (1994). *Institutionen ökonomisch analysiert: Zur jüngeren Entwicklung auf einem Gebiet der Wirtschaftstheorie* (Vol. 1786, Uni-Taschenbücher). Tübingen: Mohr.

Roiger, M. B. (2007). *Gestaltung von Anreizsystemen und Unternehmensethik: Eine norm- und wertbezogene Analyse der normativen Principal-Agent-Theorie*. Wiesbaden: Deutscher Universitäts-Verlag.

Sanderson, I. (2000). Evaluation in complex policy systems. *Evaluation*, *6*(4), 433–454. https://doi.org/10.1177/13563890022209415

Sanderson, I. (2002). Evaluation, policy learning and evidence-based policy making. *Public Administration*, *80*(1), 1–22. https://doi.org/10.1111/1467-9299.00292

Scheirer, M. A. (1998). Commentary: Evaluation planning is the heart of the matter. *American Journal of Evaluation*, *19*(3), 385–391.

Schnell, R., Hill, P. B., & Esser, E. (2018). *Methoden der empirischen Sozialforschung.* (11., überarbeitete Auflage). Berlin: De Gruyter Oldenbourg.

Sedlmeier, P., & Renkewitz, F. (2018). *Forschungsmethoden und Statistik für Psychologen und Sozialwissenschaftler* (3., aktualisierte und erweiterte Auflage). Hallbergmoos: Pearson.

Stockmann, R., Meyer, W., & Schenke, H. (2011). Unabhängigkeit von Evaluationen. *Zeitschrift für Evaluation*, *10*(1), 39–67.

Stufflebeam, D. L., & Coryn, C. (2014). *Evaluation theory, models, and applications* (2. Auflage, Vol. 50, Research methods for the social sciences). San Francisco, CA: Jossey Bass Ltd.

Stufflebeam, D. L. (1994). Empowerment evaluation, objectivist evaluation, and evaluation standards: Where the future of evaluation should not go and where it needs to go. *Evaluation Practice, 15*(3), 321–338. https://doi.org/10.1177/109821409401500313

THE LSE GV314 GROUP (2013). Evaluation under contract: Government pressure and the production of policy research. *Public Administration*, *92*(1), 224–239. https://doi.org/10.1111/padm.12055

TheWindowsClub (2015). Most commonly used email addresses and service providers. Abgerufen am 3. Januar 2019, von https://www.thewindowsclub.com/commonly-used-email-addresses.

Top Universities (2018). Top universities in the US 2019. Abgerufen am 3. Januar 2019, von https://www.topuniversities.com/university-rankings-articles/world-university-rankings/top-universities-us-2019.

Turner, D. (2003). *Evaluation ethics and quality results of a survey of Australasian Evaluation Society members*. Abgerufen am 23. März 2019, von https://www.aes.asn.au/images/stories/files/About/Documents%20-%20on-going/ethics_survey_summary.pdf.

United Nations (2019). UN system documentation. Abgerufen am 4. April 2019, von http://research.un.org/en/docs/unsystem/sa.

Universität Zürich (2019). Pearson Chi-Quadrat-Test (Kontingenzanalyse). Abgerufen am 20. Mai 2019, von https://www.methodenberatung.uzh.ch/de/datenanalyse_spss/zusammenhaenge/pearsonzush.html.

usa.gov (2019). State government. Abgerufen am 11. April 2019, von https://www.usa.gov/states-and-territories.

Valovirta, V. (2002). Evaluation utilization as argumentation. *Evaluation*, *8*(1), 60–80. https://doi.org/10.1177/1358902002008001487

Van der Knaap, P. (2004). Theory-based evaluation and learning: Possibilities and challenges. *Evaluation*, *10*(1), 16–34. https://doi.org/10.1177/1356389004042328

Van Slyke, D. M. (2006). Agents or stewards: Using theory to understand the government-nonprofit social service contracting relationship. *Journal of Public Administration Research and Theory*, *17*(2), 157–187. https://doi.org/10.1093/jopart/mul012

Waterman, R. W., & Meier, K. J. (1998). Principal-agent models: An expansion? *Journal of Public Administration Research and Theory*, *8*(2), 173–202. https://doi.org/10.1093/oxfordjournals.jpart.a024377

Webengage (2019). 28 tips to avoid spam filters when doing email marketing. Abgerufen am 4. Oktober 2019, von https://webengage.com/blog/how-to-avoid-spam-filters-when-sending-emails/.

Wenger, E., & Terberger, E. (1988). Die Beziehung zwischen Agent und Prinzipal als Baustein einer ökonomischen Theorie der Organisation. *WiSt*, *10*, 506–514.

Widmer, T. (2012). Unabhängigkeit in der Evaluation. *LeGes-Gesetzgebung & Evaluation*, 23(2), 129-147. https://doi.org/10.5167/uzh-65019

Wright, P., Mukherji, A., & Kroll, M. J. (2001). A reexamination of agency theory assumptions: extensions and extrapolations. *The Journal of Socio-Economics*, *30*(5), 413–429. https://doi.org/10.1016/S1053-5357(01)00102-0

Young, K., Ashby, D., Boaz, A., & Grayson, L. (2002). Social science and the evidence-based policy movement. *Social Policy and Society*, *1*(3), 215–224. https://doi.org/10.1017/S1474746402003068

Zaggl, M. A. (2012). *Cooperation and reciprocity in two-sided principal-agent relations: An evolutionary perspective* (Dissertation). Hamburg-Harburg: Technische Universität Hamburg-Harburg.

Anhangsverzeichnis

Anhang

A. Variablenübersicht

Tabelle Anhang 1: Zusammenfassende Statistik und Operationalisierung der Variablen

	Variable	Zusammenfassende Statistik	Operationalisierung
Abhängige Variablen	*Destruktive Beeinflussungsart*	*Anteile (N):* Nie 15.4 (4) Einmal 65.4 (17) Mehr als einmal 19.2 (5) (Gesamt N: 26; Fehlend: 36)	Additiver Index aus Q6.4: 0 = Nein, das habe ich nie gemacht 1 = Ja, das habe ich einmal gemacht 2 = Ja, das habe ich mehr als einmal gemacht 3 = Ja, das habe ich oft gemacht
	Konstruktive Beeinflussungsart	*Anteile (N):* Mehr als einmal 39.1 (9) Oft 60.9 (14) (Gesamt N: 23; Fehlend: 39)	Additiver Index aus Q6.3: 0 = Nein, das habe ich nie gemacht 1 = Ja, das habe ich einmal gemacht 2 = Ja, das habe ich mehr als einmal gemacht 3 = Ja, das habe ich oft gemacht
	Anreizsystem	*Anteile (N):* Nie 55.6 (10) Einmal 38.9 (7) Mehr als einmal 5.6 (1) (Gesamt N: 18; Fehlend: 44)	Additiver Index aus Q5.5: 0 = Nein, das habe ich nie gemacht 1 = Ja, das habe ich einmal gemacht 2 = Ja, das habe ich mehr als einmal gemacht 3 = Ja, das habe ich oft gemacht
	Direkter Einfluss	*Anteile (N):* Nie 72.2 (13) Einmal 16.7 (3) Mehr als einmal 11.1 (2) (Gesamt N: 18; Fehlend: 44)	Additiver Index aus Q5.5_2, Q5.5_4, Q5.5_6: 0 = Nein, das habe ich nie gemacht 1 = Ja, das habe ich einmal gemacht 2 = Ja, das habe ich mehr als einmal gemacht 3 = Ja, das habe ich oft gemacht
	Erwartung an Unabhängigkeit	*Anteile (N):* Eher tief 6.3 (2) Eher hoch 43.8 (14) Hoch 31.3 (10) Sehr hoch 18.8 (6) (Gesamt N: 32; Fehlend: 30)	Umpolung Items Q3.10a_1, Q3.10a_3, Q3.10b_5, Q3.10b_6, Additiver Index aus Q3.10a/b: 0 = Gar nicht 1 = Tief 2 = Eher tief 3 = Eher hoch 4 = Hoch 5 = Sehr hoch
Unabhängige Variablen	*Konfliktverhältnis*	*Anteile (N):* 0 Überhaupt nicht konfliktgeprägt: 11.1 (4) 1: 16.7 (6) 2: 22.2 (8) 3: 25.0 (9) 4: 2.8 (1) 5: 11.1 (4) 6: 11.1 (4) (Gesamt N: 36; Fehlend: 26)	Kategoriale Variable: Skala von 0 (= überhaupt nicht konfliktgeprägt) bis 10 (= äusserst konfliktgeprägt)
	Konflikthäufigkeit	*Anteile (N):* Nie 17.6 (3) Eher selten 23.5 (4) Weder noch 17.6 (3) Eher häufig 17.6 (3) Sehr häufig 17.6 (3) Äusserst häufig 5.9 (1) (Gesamt N: 17; Fehlend: 45)	Additiver Index aus Q4.4 (Konfliktgründe): 0 = Nie 1 = Sehr selten 2 = Eher selten 3 = Weder noch 4 = Eher häufig 5 = Sehr häufig

			6 = Äusserst häufig
	Unzufriedenheit	*Anteile (N):* Sehr zufrieden 5.3 (1) Ziemlich zufrieden 31.6 (6) Mittelmässig zufrieden 15.8 (3) Ziemlich unzufrieden 31.6 (6) Sehr unzufrieden 15.6 (3) (Gesamt N: 19; Fehlend: 43)	Additiver Index aus Q5.4: 0 = Sehr zufrieden 1 = Ziemlich zufrieden 2 = Mittelmässig zufrieden 3 = Ziemlich unzufrieden 4 = Sehr unzufrieden
	Schwierigkeiten	*Anteile (N):* Nie 10 (3) Sehr selten 30 (9) Eher selten 26.7 (8) Weder noch 13.3 (4) Eher häufig 6.7 (2) Sehr häufig 13.3 (4) (Gesamt N: 30; Fehlend: 32)	Additiver Index aus Q4.5: 0 = Nie 1 = Sehr selten 2 = Eher selten 3 = Weder noch 4 = Eher häufig 5 = Sehr häufig 6 = äusserst häufig
	Vertrautheit	*Anteile (N):* Eher schlecht vertraut 22.2 (2) Eher gut vertraut 66.7 (6) Sehr gut vertraut 11.1 (1) (Gesamt N: 9; Fehlend: 53)	Kategoriale Variable: 0 = Sehr schlecht vertraut 1 = Eher schlecht vertraut 2 = Eher gut vertraut 3 = Sehr gut vertraut
	Berufserfahrung	*Anteile (N):* Äusserst unerfahren 9.3 (5) Sehr unerfahren 22.2 (12) Eher unerfahren 24.1 (13) Weder noch 18.5 (10) Eher erfahren 14.8 (8) Sehr erfahren 5.6 (3) Äusserst erfahren 5.6 (3) (Gesamt N: 54; Fehlend: 8)	Additiver Index aus Q3.2 und Q3.3: 0 = Äusserst unerfahren 1 = Sehr unerfahren 2 = Eher unerfahren 3 = Weder noch 4 = Eher erfahren 5 = Sehr erfahren 6 = Äusserst erfahren
Kontrollvariablen	*Destruktive Beeinflussungsart Allgemein*	*Anteile (N):* Nie 48.1 (13) Einmal 33.3 (9) Mehr als einmal 14.8 (4) Oft 3.7 (1) (Gesamt N: 27; Fehlend: 35)	Additiver Index aus Q6.6_2 und Q6.6_4: 0 = Nein, das habe ich nie gemacht 1 = Ja, das habe ich einmal gemacht 2 = Ja, das habe ich mehr als einmal gemacht 3 = Ja, das habe ich oft gemacht
	Konstruktive Beeinflussungsart Allgemein	*Anteile (N):* Einmal 10.3 (3) Mehr als einmal 31.0 (9) Oft 58.6 (17) (Gesamt N: 29; Fehlend: 33)	Additiver Index aus Q6.6_1 und Q6.6_3: 0 = Nein, das habe ich nie gemacht 1 = Ja, das habe ich einmal gemacht 2 = Ja, das habe ich mehr als einmal gemacht 3 = Ja, das habe ich oft gemacht
	Beeinflussungsintention	*Anteile (N):* Wenig destruktiv 69.2 (18) Mittelmässig destruktiv 26.9 (7) Sehr destruktiv 3.8 (1) (Gesamt N: 26; Fehlend: 36)	Additiver Index aus Q6.5: 0 = Konstruktiv 1 = Wenig destruktiv 2 = Mittelmässig destruktiv 3 = Ziemlich destruktiv 4 = Sehr destruktiv
	Soziale Erwünschtheit	*Anteile (N):* Trifft eher nicht zu 25.0 (7) Trifft eher zu 71.4 (20) Trifft genau zu 3.6 (1) (Gesamt N: 28; Fehlend: 34)	Kategoriale Variable: 1 = Trifft nicht zu 2 = Trifft eher nicht zu 3 = Trifft eher zu 4 = Trifft genau zu

Soziodemographische Variablen	*Alter*	*Mittelwert:* 46.30 SD: 12.763 Min.: 29 Max.: 77 (Gesamt N: 30; Fehlend: 32)	Alter in Jahre
	Bildung	*Anteile (N):* College 3.2 (1) Bachelorabschluss 6.5 (2) Masterabschluss 51.6 (16) Berufsabschluss 6.5 (2) Doktortitel 32.3 (10) (Gesamt N: 31; Fehlend: 31)	Kategoriale Variable: 0 = High School 1 = College 2 = Associate Degree 3 = Bachelorabschluss 4 = Masterabschluss 5 = Berufsabschluss 6 = Doktortitel
	Geschlecht	*Anteile (N):* Weiblich 61.3 (19) Männlich 38.7 (12)	Dummy: 1 = Weiblich 2 = Männlich
Ergänzende Variablen	*Evaluationsjahre*	*Anteile (N):* Weniger als 1 Jahr 10.7 (6) 1-5 Jahre 46.4 (26) 6-10 Jahre 26.8 (15) 11-15 Jahre 5.4 (3) Mehr als 15 Jahre 10.7 (6) (Gesamt N: 56; Fehlend: 6)	Kategoriale Variable: 0 = Weniger als 1 Jahr 1 = 1-5 Jahre 2 = 6-10 Jahre 3 = 11-15 Jahre 4 = Mehr als 15 Jahre
	Evaluationsanzahl	*Anteile (N):* 1-5 Evaluationen 38.9 (21) 6-20 Evaluationen 37.0 (20) Mehr als 20 Evaluationen 24.1 (13) (Gesamt N: 54; Fehlend: 8)	Kategoriale Variable: 0 = 1-5 Evaluationen 1 = 6-20 Evaluationen 2 = Mehr als 20 Evaluationen
	Auftraggeber-Sektor	*Anteile (N):* Öffentlicher Sektor 40.4 (21) Universität, Hochschule 13.5 (7) Privatsektor 7.7 (4) NGO 26.9 (14) Sonstiges 11.5 (6) (Gesamt N: 52; Fehlend: 10)	Kategoriale Variable: 0 = Öffentlicher Sektor 1 = Universität, Hochschule 2 = Privatsektor 3 = NGO 4 = Sonstiges
	Evaluator-Sektor	*Anteile (N):* Öffentlicher Sektor 14.3 (7) Universität, Hochschule 20.4 (10) Privatsektor 40.8 (20) NGO 12.2 (6) Sonstiges 12.2 (6) (Gesamt N: 49; Fehlend: 13)	Kategoriale Variable: 0 = Öffentlicher Sektor 1 = Universität, Hochschule 2 = Privatsektor 3 = NGO 4 = Sonstiges
	Berufsfeld	*Anteile (N):* Architektur und Engineering 3.2 (1) Business und Finance 9.7 (3) Öffentlicher Sektor 22.6 (7) Gesundheitswesen 22.6 (7) Lebens-, Natur- und Sozialwissenschaften 16.1 (5) Sonstiges 25.8 (8) (Gesamt N: 31; Fehlend: 31)	Kategoriale Variable: 0 = Tierpflege und Wissenschaft 1 = Architektur und Engineering 2 = Kunst und Design 3 = Business und Finance 4 = Computer und IT 5 = Unterhaltung und Sport 6 = Land-, Fischerei- und Forstwirtschaft 7 = Öffentlicher Sektor 8 = Gesundheitswesen 9 = Lebens-, Natur- und Sozialwissenschaften 10 = Sonstiges

Evaluationsart	*Anteile (N):* Externe Evaluationen 83.3 (40) Interne Evaluationen 16.7 (8) (Gesamt N: 48; Fehlend: 14)	Dummy: 0 = Externe Evaluationen 1 = Interne Evaluationen
Anteil Direktvergabe	*Mittelwert:*49.67 SD: 38.166 Min.: 0 Max.: 100 (Gesamt N: 42; Fehlend: 20)	Direktvergabe in Prozent
Wichtigkeit Vergabekriterien	*Anteile (N):* Unwichtig 2.6 (1) Eher wichtig 5.3 (2) Wichtig 23.7 (9) Sehr wichtig 68.4 (26) (Gesamt N: 38; Fehlend: 24)	Additiver Index aus Q3.9: 0 = Überhaupt nicht wichtig 1 = Unwichtig 2 = Eher unwichtig 3 = Eher wichtig 4 = Wichtig 5 = Sehr wichtig
Unzufriedenheit Vergangenheit	*Anteile (N):* Nein 42.9 (15) Ja 57.1 (20) (Gesamt N: 35; Fehlend: 27)	Dummy: 0 = Nein 1 = Ja
Konflikt	*Anteile (N):* Nein 48.65 (18) Ja 48.65 (18) Weiss nicht 2.7 (1) (Gesamt N: 37; Fehlend: 25)	Dummy: 0 = Nein 1 = Ja -88 = Weiss nicht
Änderungen	*Anteile (N):* Nein 12.6 (5) Ja 71.9 (23) Weiss nicht 12.5 (4) (Gesamt N: 32; Fehlend: 30)	Dummy: 0 = Nein 1 = Ja -88 = Weiss nicht
Reaktion	*Anteile (N):* Nicht vorgenommen 3.3 (1) Vorgenommen 26.7 (8) Kompromiss 70.0 (21) (Gesamt: 30; Fehlend: 32)	Kategoriale Variable: 0 = Die Änderungen wurden nicht vorgenommen 1 = Die Änderungen wurden vorgenommen 2 = Ein Kompromiss wurde gefunden
Unterstellung	*Anteile (N):* Keine Unterstellung 94 (30) Weiss nicht 6 (2) (Gesamt:32; Fehlend 30)	Kategoriale Variable: 0 = Nein, dies wurde mir noch nie unterstellt 1 = Ja, dies wurde mir einmal unterstellt 2 = Ja, dies wurde mir mehr als einmal unterstellt 3 = Ja, dies wurde mir oft unterstellt -88 = Weiss nicht
Wahrnehmung Einflussstärke	*Anteile (N):* 0 Überhaupt nicht stark: 6.7 (2) 1: 10.0 (3) 2: 10.0 (3) 3: 6.7 (2) 4: 10.0 (3) 5: 23.3 (7) 6: 3.3 (1) 7: 10.0 (3) 8: 16.7 (5) 9: 3.3 (1) (Gesamt N: 30; Fehlend: 32)	Kategoriale Variable: Skala von 0 (= überhaupt nicht stark) bis 10 (= äusserst stark)

Wahrnehmung Unabhängigkeit	*Anteile (N):* 3: 23.3 (7) 4: 33.3 (10) 5 Sehr unabhängig: 43.3 (13) (Gesamt N: 30; Fehlend: 32)	Kategoriale Variable: Skala von 0 (= überhaupt nicht unabhängig) bis 5 (= sehr unabhängig)
Wichtigkeit Unabhängigkeit	*Anteile (N):* 1: 3.2 (1) 2: 9.7 (3) 3: 9.7 (3) 4: 22.6 (7) 5 Sehr wichtig 22.6 (7) (Gesamt N: 31; Fehlend: 31)	Kategoriale Variable: Skala von 0 (= überhaupt nicht wichtig) bis 5 (= sehr wichtig)
Standardkenntnis	*Anteile (N):* Nein 70.0 (21) Ja 30.0 (9) (Gesamt N: 30; Fehlend: 32)	Dummy: 0 = Nein 1 = Ja
Standardwichtigkeit	*Anteile (N):* Eher unwichtig 11.1 (1) Eher wichtig 77.8 (7) Sehr wichtig 11.1 (1) (Gesamt N: 9; Fehlend: 53)	Kategoriale Variable: 0 = Sehr unwichtig 1 = Eher unwichtig 2 = Eher wichtig 3 = Sehr wichtig
Prävention	*Anteile (N):* Meta-Evaluation 5.0 (6) Interne / externe Meldestelle 3.4 (4) Engere Zusammenarbeit 15.1 (18) Gegenseitiges Verständnis 22.7 (27) Diskussion negativer Resultate 10.9 (13) Betonung der Verantwortung 8.4 (10) Verbesserte Datendokumentation 9.2 (11) Formelles Evaluationsprotokoll 12.6 (15) Neutrale Intervention Dritte 5.0 (6) Hinweis Stellungnahme 5.0 (6) Sonstiges 2.5 (3)	Kategoriale Variable mit Mehrfachantworten 0 = Meta-Evaluation 1 = Interne / externe Meldestelle 2 = Engere Zusammenarbeit 3 = Gegenseitiges Verständnis 4 = Diskussion negativer Resultate 5 = Betonung der Verantwortung 6 = Verbesserte Datendokumentation 7 = Formelles Evaluationsprotokoll 8 = Neutrale Intervention Dritte 9 = Hinweis Stellungnahme 10 = Sonstiges

(Quelle: eigene Darstellung)

B. Externer Anhang

a. Codebuch

b. SPSS Datensatz

c. SPSS Syntaxdateien